

UNCONFINED COMPRESSIVE STRENGTH PREDICTION USING DRILLING
PARAMETERS AND ANALYZING FEATURE IMPORTANCE THROUGH
PRINCIPAL COMPONENTS ANALYSIS

by

Muhammed Esad Kaya

© Copyright by Muhammed Esad Kaya, 2022

All Rights Reserved

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Petroleum Engineering).

Golden, Colorado

Date _____

Signed: _____
Muhammed Esad Kaya

Signed: _____
Dr. Alfred William Eustes III
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Jennifer Miskimins
Professor and Head
Department of Petroleum Engineering

ABSTRACT

Knowledge of geomechanical properties is beneficial if not essential for drilling and completion operations in the oil and gas industry. The Unconfined Compressive Strength (UCS) is the maximum compressive force applied to cylindrical rock samples without breaking under unconfined conditions. Unconfined Compressive Strength (UCS) is one of the key criteria to ensure safe, efficient, and successful drilling operations, and estimation of UCS is vital to avoid wellbore stability problems that are inversely correlated with the pace of drilling operations. Furthermore, UCS is an essential input to ensure the success of completion operations such as acidizing and fracturing.

Different methods are available to estimate UCS. The common practice to estimate UCS is to conduct experiments with a laboratory testing setup. These laboratory experiments are considered the most accurate way to measure UCS, but they are destructive, time-consuming, and expensive. Alternatively, empirical equations are derived to estimate UCS from well-logging tool readings. These empirical equations are generally derived from physical properties such as interval transit time, porosity, and Young's modulus. However, most of these equations are not generic, and their applicability for other formation types is limited.

The limitations of existing methods to estimate UCS promoted the development of data-driven solutions to estimate UCS. The data-driven methods include but are not limited to basic regression, machine learning, and deep learning algorithms. Data-driven methods to identify patterns in the data to estimate geomechanical parameters are considered to be implemented for drilling operations.

This study proposes methods to assist safe and successful drilling operations while eliminating the need for coring, saving a vast amount of time and money by estimating UCS from drilling parameters instantaneously. The goal is to develop a machine-learning algorithm to analyze and process high-frequency data to estimate UCS instantaneously while drilling, allowing safer and more efficient drilling operations.

The drilling data used to train, validate, and test the machine learning model is re-purposed from data collected during drilling in a previous study. The algorithm consists of a data processing method called Principal Component Analysis (PCA) to indicate the importance of each parameter by quantifying their variance contribution. Random Forest machine learning algorithm is utilized

to build a regression model to estimate UCS. The regression model developed uses Depth, Rotation per Minute (RPM), Weight-on-Bit (WOB), Torque, Rate of Penetration (ROP), Mechanical Specific Energy (MSE), and Normalized Formation Penetration Index (N-FPI) as an input to estimate UCS.

The blind data split from the original data set is used to confirm the veracity and applicability of the algorithm. The tests conducted on the final model indicated that the algorithm built is flexible enough to adjust to different conditions and formations to deliver accurate estimations. With a proper allocation of computational power and high-quality drilling data, the algorithm built can be trained to estimate UCS instantaneously while drilling.

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
ACKNOWLEDGMENTS	xiii
CHAPTER 1 INTRODUCTION	1
1.1 Motivation	3
1.2 Objectives	3
1.3 Thesis Organization	3
CHAPTER 2 OVERVIEW	5
2.1 The Unconfined Compressive Strength	5
2.2 Principal Component Analysis	13
2.3 Tree-Based Algorithms	17
2.3.1 Random Forest	18
CHAPTER 3 BACKGROUND INFORMATION ABOUT DATASET AND PCA	23
3.1 Background Information about the Dataset	23
3.1.1 Drilling Depth	25
3.1.2 Weight on Bit (WOB)	26
3.1.3 Rate of Penetration (ROP)	26
3.1.4 Normalized Field Penetration Index (N-FPI)	26
3.1.5 Mechanical Specific Energy (MSE)	27

3.2	Uniqueness with Dataset and Challenges	27
3.3	Principal Component Analysis	28
3.3.1	The Concept of PCA	29
3.3.1.1	Eigenvalue Decomposition (Matrix Diagonalization)	31
3.3.1.2	Singular Value Decomposition	32
3.3.2	The Results of PCA	33
3.3.2.1	Explained Variance Analysis	33
3.3.2.2	Feature Importance Analysis	35
CHAPTER 4 MODELING AND TESTING		37
4.1	Machine Learning Methods	37
4.1.1	Regression Models	38
4.1.1.1	Linear Regression	38
4.1.1.2	Polynomial Regression	40
4.1.1.3	Support Vector Regression	41
4.1.1.4	Random Forest Regression	42
4.2	Possible Problems and Solutions	43
4.2.1	Bias and Variance	44
4.2.2	Underfitting and Overfitting	44
4.2.3	Bias and Variance Trade-off	45
4.3	Hyper Parameter Tuning	46
4.4	Evaluating Performance of the Models	48
4.5	Final Architecture of the Model	49
4.6	Discussion on Challenges and Changes	51
4.7	Potential Implementation of the Model for Field Applications	55

CHAPTER 5 RESULTS AND TECHNICAL EVALUATION	57
5.1 The development phase of the algorithm	57
5.1.1 Initial Results	57
5.1.2 Models Fitted without PCA	60
5.1.3 Multi-output Random Forest Model	62
CHAPTER 6 SUMMARY AND CONCLUSIONS	66
CHAPTER 7 FUTURE WORK	69
REFERENCES CITED	70
APPENDIX ASUPPLEMENTAL FILE	75
APPENDIX BCOPYRIGHT PERMISSIONS	76

LIST OF FIGURES

Figure 2.1	Stress distribution of different test methods. (A) Uniaxial compression; (B) Uniaxial strain; (C) Tensile; (D) Hydrostatic; (E) Triaxial compression; (F) Polyaxial compression ,©Reprinted from The Applied Petroleum Geomechanics, Vol. I, Zhang,J,J, Chapter III, 86, Copyright (2020), with permission from Elsevier.	6
Figure 3.1	Three Concrete Blocks while curing and demolding.©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi	24
Figure 3.2	Experimental Setup ©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi	25
Figure 3.3	The architecture of algorithm built by ©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi	28
Figure 3.4	Histogram of UCS - Complete Dataset	29
Figure 3.5	Explained Variance within the Training Dataset	34
Figure 3.6	Explained Variance within Iris Dataset	34
Figure 3.7	Feature Matrix indicating influence of each variable to PCs	36
Figure 4.1	The backward elimination flowchart	40
Figure 4.2	Linear Regression, 2 nd , 300 th degree Polynomial Regression models on the same training dataset ©2019 Kiwisoft S.A.S. Published by O’Reilly Media, Inc. Used with permission	41
Figure 4.3	One-dimensional linear SVR example	42
Figure 4.4	(Left) The relation between WOB,ROP, and RPM based on split rules, (Right) The decision tree to predict ROP. Modified from Joshi 2021.	43
Figure 4.5	An Example of Overfitting the training data ©2019 Kiwisoft S.A.S. Published by O’Reilly Media, Inc. Used with permission	45
Figure 4.6	Bias - Variance Trade-off	46
Figure 4.7	K-fold Cross Validation Example	47
Figure 4.8	RandomizedSearchCV-GridSearchCV Steps of Implementation	48

Figure 4.9	The Final Architecture of prediction model.	50
Figure 4.10	The Explained Variance Retained by PCs for Datasets 1 to 4.	52
Figure 4.11	The Feature Matrix of Datasets 1 - 2.	53
Figure 4.12	The Feature Matrix of Datasets 3 - 4.	54
Figure 4.13	Potential Implementation on Field Application Workflow	55
Figure 5.1	Comparison between predicted and actual UCS with 20, 50, 100 moving averages.	60
Figure 5.2	Predicted and actual UCS data with 10000 moving average - Predicted Values (Blue), Actual Values(Black).	61
Figure 5.3	Comparison between prediction and test UCS data with 10000 moving average.	62
Figure 5.4	Target UCS (Black) vs Predicted UCS(Blue).	64
Figure 5.5	Actual and Predicted MSE with the Final Model.	64
Figure 5.6	Predicted vs Actual UCS and MSE values after applying 10,000 moving average.	65

LIST OF TABLES

Table 3.1	Properties of Cryogenic Samples. Modified from (Joshi 2021)	24
Table 3.2	Properties of Cellular Samples. Modified from (Joshi 2021)	24
Table 5.1	Initial models fitted with sub datasets after implementing PCA	58
Table 5.2	Hyper-parameters of the model fitted with four PCs	59
Table 5.3	The parameters used for Final Cases	63
Table 5.4	Hyper-parameters of the model fitted with six features and two target outputs . .	63
Table 5.5	The prediction accuracy for each model fit	63

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
ASTM	American Society for Testing and Materials
DL	Deep Learning
ED	Eigenvalue Decomposition
GBDT	Gradient Boosting Decision Tree
ICA	Independent Component Analysis
ISRM	International Society for Rock Mechanics
MAE	Mean Absolute Error
MAEP	Mean Average Error Percentage
MAPE	Mean Absolute Percentage Error
MC	Minor Components
MCA	Minor Component Analysis
MFL	Mamdani Fuzzy Logic
ML	Machine Learning
MLP	Multi-layer Perception
MS	Minor Subspace
MSE	Mechanical Specific Energy
NN	Neural Network
PC	Principal Components
PCA	Principal Component Analysis

PS Principal Subspace

RF Random Forest

RGB Red, Green, and Blue

RMSE Root Mean Square Error

ROP Rate of Penetration

RPM Rotation per Minute

SFL Sugeno Fuzzy Logic

SLM Statistical Learning Model

SVD Singular Value Decomposition

SVM Support Vector Machine

SVR Support Vector Regression

TD Torque and Drag

UCS Unconfined Compressive Strength

VAF Variance Account for

WOB Weight-on-Bit

WWSLM Windows Statistical Learning Model

XRF X-Ray Fluorescence

ACKNOWLEDGMENTS

I want to acknowledge the help of people and organizations that made this research possible by contributing to my research, studies, and daily life.

First and foremost, I would like to thank my sponsor, MTA (General Directorate of Mineral Research and Exploration of Turkey), for financially sponsoring me through my studies to achieve excellence in the exploration of Geothermal resources in Turkey.

I want to remind the memories of Dr. Tutuncu and thank her for opening the doors of CSM to me. She will always be remembered for her legacy here at CSM. I will remember her as a hard-working, deeply knowledgeable, kind, and lovely advisor.

I want to thank Dr. Eustes for supporting me through all my studies. His trust in me motivated me to complete my research. I want to thank him for being there to answer my questions and for his wisdom.

I want to express my most profound appreciation to my committee members for their help and for making this research possible.

I am honored to gain an opportunity to learn from professors and staff here at CSM. I would like to thank Dr. Eustes, Dr. Ozkan, Dr. Prasad, Prof. Crompton, and Dr. Fleckenstein for their classes.

I would like to thank all the Petroleum Engineering Department members, especially Denise and Rachel, for their support throughout my studies.

Sincere thanks to my friends, Hazar, Deep, Santiago, Mansour, Mansour Ahmed, Val, Gizem, and Roy, for supporting me through all hardiness. Thanks to you guys, I look at life from a broader perspective.

Special thanks to Nehra for her company, love, and support that motivated me to complete my studies and made my life better in every way. Thank you for being there for me.

Finally, I would like to thank my parents, Alime and Abdullah, for being my teachers throughout my life. Everything I achieved was thanks to you two.

CHAPTER 1

INTRODUCTION

Historical data has always been an essential part of the oil and gas industry. The industry has become data-intensive with the recent advancements in data collection due to more durable and reliable sensors. However, the amount of data utilized to improve the efficiency of future operations is still a fraction of the data collected. The oil and gas industry is becoming more aware of the potential uses of these data. The utilization of data is being recognized as the most efficient method for reducing cost by increasing operational efficiency and creating safer, more sustainable developments (Løken et al. 2020).

Specifically, the drilling industry has started investing more into the automation of drilling operations due to efficiency, safety, and cost concerns. In the last decade, the increasing computational powers and the digitization of rig parts have allowed the industry to utilize machine learning algorithms (ML) for most drilling operations. By implementing data-driven solutions through ML algorithms, the industry is working on building automated drilling systems that can conduct drilling operations without human input or recommend an efficient solution for the safety of operations. One objective of the drilling industry is to increase the efficiency of drilling operations by reducing capital and operational expenses with the implementation of data-driven solutions. Knowing subsurface conditions and geomechanical properties is essential to achieving this objective. Especially by gaining more knowledge about geomechanical properties, wellbore stability can be improved while drilling by avoiding hole collapse, stuck pipe, tight hole, kicks, and loss circulation.

Drilling parameters have been recognized and used as an indicator of formation parameters, and estimating geomechanical properties from drilling parameters has been stated as an important topic, with studies conducted since the late 1950s (Combs 1968; Cunningham and Eenink 1959). Early studies completed by Bourgoyne and Young (1974) showed that pore pressure could be determined from drilling parameters with 1 lb/gal standard deviation on Gulf Coast, and Majidi et al. (2017) observed similar results in estimating formation pore pressure from MSE and drilling parameters from a study with a similar intent. Some of these models for pore pressure estimation by Jorden and Shirley (1966) as well as Rehm and McClendon (1971) are still being used in

the industry; thanks to their practicality. The objective of these studies is to indicate the importance of estimating formation parameters and efficiency of the operations. Likewise, Unconfined Compressive Strength (UCS) has been known as an essential parameter as it is a key input to avoiding possible wellbore failures by implementing a robust mud weight window and deciding an aggressiveness of bit (Nabaei et al. 2010).

The Unconfined Compressive Strength is the maximum compressive strength that a cylindrical rock sample can withstand under unconfined conditions. The UCS is also known as a uniaxial compressive strength because the compressive stress is applied along only one axis while measuring the rock strength. The impact of rock hardness, also known as rock strength, on drilling performance has always been an important issue and has been investigated since the early 1960s (Spaar et al. 1995). In addition, unconfined compressive strength is one of the essential parameters when deciding on bit aggressiveness (Spaar et al. 1995).

The early studies indicated a strong correlation between rock hardness and drilling performance, and it is also observed that other drilling parameters such as weight on bit, revolution per minute, and bit type are required to predict the drilling efficiency among rock hardness measurements (Gstalter and Raynal 1966). The study indicates that estimation or measurement of UCS is essential to avoid wellbore stability problems while drilling. In addition, a study conducted by Spaar et al. (1995) shows a strong correlation between the formation drillability with UCS and friction angle as these parameters are essential for bit selection and the selection of appropriate aggressiveness for a bit can improve overall drilling performance substantially.

The empirical equations derived from well logging tool readings and rock strength tests run in laboratory conditions are the most common methods to estimate UCS. However, data-driven solutions to estimate these parameters are becoming more common as these methods are getting more robust thanks to studies conducted to observe their veracity and versatility with available geomechanical and drilling data. Also, an exponential increase in the number of drilling operations conducted in unconventional reservoirs brought the need for a more sophisticated and cheaper method to estimate geomechanical parameters as these reservoirs commonly have non-linear behavior and coring in a horizontal section of the wells drilled through the unconventional reservoirs is harder to conduct. These reasons indicate a need for a faster and cheaper method to estimate geomechanical parameters.

1.1 Motivation

This thesis proposes to build a data-driven solution to estimate UCS instantaneously from drilling parameters by utilizing Random Forest regression model. The reviewed studies conducted by a vast amount of scientists indicate that the data-driven methods can improve the efficiency of operations in various ways by introducing sophisticated solutions such as predicting ROP (Hegde et al. 2015), estimating drilling optimization parameters (Nasir and Rickabaugh 2018), indicating the development of dominant water channels (Chen et al. 2019), predicting casing failures (Song and Zhou 2019), and predicting possible drilling incidents (AlSaihati et al. 2021). This study is solely motivated to provide key input parameter to avoid potential wellbore stability problems and drilling accidents by indicating rock strength changes within the formation. Furthermore, the final model from this study can be adjusted and developed to integrate into a fully automated drilling system as instantaneous UCS pattern indicator, which will be one of the essential steps for the next level of drilling automation.

1.2 Objectives

The main objectives of this thesis are:

- Develop a machine-learning model that can be trained to estimate unconfined compressive strength from drilling parameters instantaneously.
- Provide changes in Unconfined Compressive Strength based on drilling parameters instantaneously to avoid possible wellbore failures and drilling accidents.
- Utilize principal component analysis to analyze feature importance using available drilling data.
- Study the implementation of Random Forest regression algorithm to build a robust regression model to estimate certain geomechanical parameters (i.e., UCS, and Mechanical Specific Energy).

1.3 Thesis Organization

This thesis consists of six chapters. The summary of each chapter is presented as follows:

- Chapter 1 introduces the background, including the objectives of this study, and provides the organization of the thesis.
- Chapter 2 describes the technical background and summarizes the early and most recent studies on UCS prediction, principal component analysis, and random forest algorithm.
- Chapter 3 presents background information about the data set and the results of principal component analysis.
- Chapter 4 presents Machine Learning algorithm utilized to complete the objectives of this research, includes a discussion on changes in the study based on findings, and discusses the potential implementation of the model for real-life field applications.
- Chapter 5 discusses the results and the performance of the model on the prediction of UCS.
- Chapter 6 concludes this research and states the contributions of this study.
- Chapter 7 includes suggestions for future work as a continuation of this study.

CHAPTER 2

OVERVIEW

2.1 The Unconfined Compressive Strength

The unconfined compressive strength can be defined as the maximum amount of force a cylindrical rock sample can withstand under unconfined conditions. The most common methods to estimate UCS are laboratory experiments and empirical equations derived from well logging tool readings. The laboratory experiments are conducted by using a testing setup that measures the maximum stress a sample can withstand; this method is referred to as a direct method to measure UCS, and the empirical equations derived from well-logging tool readings are referred to as an indirect method to estimate UCS (Ceryan et al. 2013). The American Society for Testing and Materials (ASTM) and the International Society for Rock Mechanics (ISRM) standardized the laboratory testing procedures that should be followed to estimate UCS. The laboratory testing procedures can vary with respect to stress distribution created around the sample. These laboratory testing methods use compressive, tensile, and triaxial stress distributions for different scenarios. The most common test method to measure the rock strength is called the uniaxial compressive test method, which is determined by applying compressive stress to the sample vertically until the sample fails (Brook 1993). The stress distributions for different test methods are shown in Figure 2.1.

The most common limitation while conducting these laboratory experiments is the quality of the core sample, as the broken or chipped sample will result in a change in stress distribution within the sample. The standardization of core sample preparation and test procedures by ASTM includes a detailed description to ensure accurate measurements. For example, according to American Society for Testing and Materials (2014), the minimum diameter of the test sample should be approximately 47-mm, and the length to diameter ratio of the test sample should be between 2.0:1 and 2.5:1 to satisfy the criterion in the majority of cases.

As mentioned before, in addition to laboratory experiments, several empirical equations were derived to estimate the rock strength, as coring operation required to obtain rock samples for these experiments is time-consuming and expensive. The empirical equations can vary to the parameter

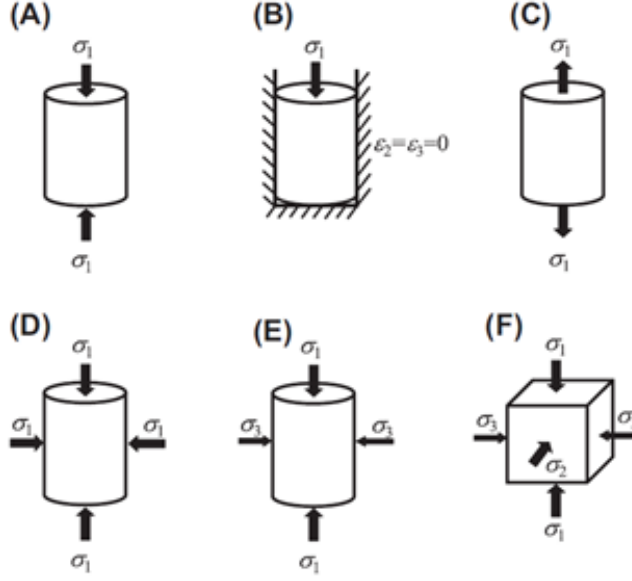


Figure 2.1: Stress distribution of different test methods. (A) Uniaxial compression; (B) Uniaxial strain; (C) Tensile; (D) Hydrostatic; (E) Triaxial compression; (F) Polyaxial compression (Zhang 2019), ©Reprinted from The Applied Petroleum Geomechanics, Vol. I, Zhang, J, J, Chapter III, 86, Copyright (2020), with permission from Elsevier.

that is planned to be estimated, and they are commonly derived for a specific formation or a section of the formation. Generally, these equations are derived from well logging tool readings such as sonic velocity, porosity, or Young's Modulus. For example, Equations 2.1 , 2.2, and 2.3 are derived using sonic velocity readings to estimate the rock strength of high porosity Tertiary shales in the Gulf of Mexico (Zhang 2019).

$$c = 5 * (V_p - 1) / \sqrt{V_p} \quad (2.1)$$

$$\sin\varphi = (V_p - 1) / (V_p + 1) \quad (2.2)$$

$$UCS = 10(V_p - 1) = 10(304.8/\Delta t - 1) \quad (2.3)$$

Where the uniaxial compressive strength, aka unconfined compressive strength (UCS), and the cohesion (c) are in MPa; the angle of friction (φ) is in degrees; the compressional sonic velocity (V_p) is in km/s; the transit time (Δt) is in $\mu s/ft$. The limitation of these empirical equations is lack of generalization as they are unique for each formation, and also, problems like low accuracy occur if a wide range of data are used to derive them.

Most laboratory testing methods used to estimate the UCS are accurate if the test sample fits within the predetermined standards, but these methods are destructive and time-consuming. On the other hand, the applicability of empirical equations to estimate UCS is limited. The empirical equations could be a better option if the time is limited. However, the accuracy of empirical equations is as good as the accuracy of the well-logging tool readings, which can introduce inaccuracy to the estimations. The limitations of laboratory testing methods and empirical correlations motivated the industry to develop an additional method to estimate UCS. With this motivation, the oil and gas industry has started to conduct studies to implement data-driven solutions as a suitable option, among other methods, to estimate parameters that require significant time and budget to measure (i.e., UCS). Especially, gaining even a limited amount of knowledge about UCS became necessary as UCS changes could help indicate potential wellbore stability problems in advance. The wellbore stability and drilling problems cost the industry a vast amount of money every year since wellbore stability problems can lead to stuck pipes, stuck tools due to differential sticking, and excessive mud losses due to tensile fractures. Some of these wellbore stability problems, such as tensile fractures and a breakout, can occur due to improper mud weight window (i.e., excessive mud weight, low mud weight). Excessive mud weight can lead to tensile fractures, which will cause loss circulation and increase the chances of differential sticking, whereas stress around the wellbore will cause breakout if the mud weight can't withstand the compressive strength of the rock (Al-Wardy and Urdaneta 2010).

The study conducted by Al-Wardy and Urdaneta (2010) indicated that the time required to deliver a well located in North Oman could be reduced from 36.8 days in 2009 to an average of 30.1 days in 2010 by understanding the geomechanics of the field. A better understanding of geomechanics is achieved by building a geomechanical model of the particular area and completing wellbore stability analysis by using vertical stress values from density logs, elastic properties and rock strength through DSI logs, minimum and maximum horizontal stress values from Minfrac/XLOT data, and stress orientation from BHI image logs. In addition to fracture tests and data collected from logs, the available formation and drilling parameters such as pore pressure, mud weights, drilling reports, and wellbore trajectory are used to complete the geomechanical model and wellbore stability analysis. This study showed that understanding geomechanics of even a particular area of a field can help reduce wellbore stability problems considerably and improve overall

drilling performance. Also, in the study, it is stated that there is a sensitive correlation between the wellbore stability problems with UCS and minimum horizontal stress.

A similar study was conducted by Klimentos (2005) to optimize drilling performance by providing optimum drilling parameters and to estimate pore pressure values and wellbore stability plan through a geomechanical model. The study focuses on optimizing drilling performance in deep-water wells, especially while drilling shaly formation. To achieve the objective of providing optimum drilling parameters, the initial Mechanical Earth Model (MEM) is developed by using well logging tool readings, mud-logs, and drilling information. Then, a proper match between logs is completed to indicate the lithology and porosity of sections. Later, the overburden stress is estimated by integrating density log readings with the MEM. The exponential extrapolation model is used to estimate the missing values for the sections where density log readings were missing to study the overall rock strength through in-situ stresses from overburden and pore pressure. After combining previous efforts with estimated pore pressure of shaly formation through compaction theory and pressure/sonic log readings, the final MEM is completed. It is indicated that the determination of MEM and using optimum mud weight windows minimized washouts and loss circulation, and it allowed them to understand better the necessity of using casing strings. The results indicated that drilling performance could be improved by understanding the geomechanics of formation as the number of days to drill and construct a well was reduced by 15 days, and \$4+ million was saved on total drilling cost.

The studies on the importance of gaining knowledge about UCS on design phases of drilling and completion operations are conducted by Brehm et al. (2006) and Al-Awad (2012). Brehm et al. (2006) completed a case study on Shenzi Field regarding the anisotropic behavior of the formations and the impacts on wellbore stability. Then, Al-Awad (2012) conducted a study that focused on the simple correlation between UCS and apparent cohesion, and throughout the study, impacts of wellbore stability issues as a result of lack of accuracy on rock strength estimation are included. Furthermore, comparison studies are conducted to question the accuracy of empirical equations to prove the versatility and veracity of these solutions. The research conducted by Meulenkamp and Alvarez (1999) compared the performance of empirical equations and used Machine Learning (ML) algorithms to estimate UCS of different rock samples to indicate if ML can be a valuable tool for the industry. Then, Chang et al. (2006) conducted a study on the accuracy of empirical equations using

a vast range of data and studied the applicability of 31 different empirical equations. Yurdakul et al. (2011) also conducted a similar study comparing the accuracy of the simple regression model and Artificial Neural Network (ANN) model in estimating UCS of sedimentary rock samples from 17 different regions of Turkey.

Furthermore, Barzegar et al. (2016) expanded the coverage of the study and compared the performance of different Machine Learning (ML) and Deep Learning (DL) models. Later, Negara et al. (2017) introduced elemental spectroscopy to the prediction model and searched for the potential impact of grain size on UCS by using the ML and DL model. The reviewed studies showed that data-driven solutions are becoming more robust.

Brehm et al. (2006) completed a case study on the wellbore located at the Shenzi Field in Green Canyon blocks 653 and 654 regarding the wellbore stability issues. The study focused on building a complex geomechanical model for wellbores where the main problems are anisotropic failure and lost circulation. The study indicates the importance of using complex and comprehensive geomechanical models while drilling the wellbore with weak rocks and overpressured zones. This combination could limit the available mud weight window, and cause wellbore stability problems if the geomechanical model does not explain these phenomena in detail. The study also states that basic geomechanical modeling, built using the earth's mechanical properties and in-situ earth's effective stresses of the region, brings an overgeneralized approach to wellbore reaction while drilling a formation where anisotropic failure occurs as a consequence of weakly bedded rocks. It is also stated that these basic geomechanical models can be improved to understand wellbore failures better if it is applied correctly by building the model with accurate data. The complexity of these models is directly correlated with the accuracy quantified of the mechanical properties (i.e., pore pressure, in-situ stress magnitudes, stress orientation, rock strength). Further discussions showed that these wellbore stability problems could be avoided if the mud weight used while drilling at Shenzi was updated based on the anisotropic behavior of shale reservoirs. The model built includes changes in in-situ stress magnitudes, stress orientation, and UCS estimations. The results indicated that the significant wellbore stability and lost circulation problems in previous exploration operations are substantially reduced and turned into manageable drilling problems.

Al-Awad (2012) conducted research focusing on the correlation between UCS and the apparent cohesion of rocks. The study also points out how important it is to know rock strength, aka UCS,

before designing drilling and completion operations to avoid possible wellbore stability problems such as sloughing shale, stuck drill pipe, tensile fractures, and breakouts. Also, in the study, it is mentioned that possible wellbore stability problems in producing wells such as sand production, perforation instability, subsidence, mechanical damage and how these problems can be foreseen in the design stage if rock strength is known. The correlation model between apparent cohesion and UCS is developed using available data from 300 different rock samples. The results showed that a simple correlation between rock apparent cohesion and UCS can be developed, and the correlation can estimate the rock mechanical properties with a 10% average error.

The research study conducted by Meulenkamp and Alvarez (1999) compared the accuracy of estimation of UCS values by utilizing regression techniques, aka empirical correlations, and Neural Network. In the study, the data set contains records of 194 rock samples ranging from weak sandstones to very strong granodiorites, and the Equotip hardness tester is used as an index test for rock strength properties. The comparison was completed between three different methods: curve fitting, multivariate regression, and NN algorithm. The coefficient of determination (R^2) of NN algorithm trained by this data set is determined as 0.967, while R^2 measured 0.957 for Multivariate Regression relation and 0.910 for curve fitting relation. The results indicated that 0.967 of actual UCS values stay within the regression line fit by NN. Even though R^2 values for NN and Multivariate Regression relation are similar, the results would possibly change if a more extensive data set is used. Also, in the study, it is observed that the statistical relations underestimated high UCS values and overestimated low UCS values as the statistical relations were based on the mean of all predictions. Even though ML has limitations, the high accuracy of predictions indicates that it is possible to develop algorithms and implement them for field applications. Furthermore, the study indicated that ML algorithms reduce the cost and time to derive empirical correlations or conduct destructive experiments.

Chang et al. (2006) reviewed and summarized the empirical equations that are derived to estimate the unconfined compressive strength and internal friction angle of sedimentary rock (shale, limestone, dolomite, and sandstone) by using physical properties (such as porosity, velocity, and modulus). The author describes the importance of deriving efficient empirical correlations by pointing out the difficulty of retrieving core samples of overburden formations, where wellbore instability problems generally occur. In the study, overall, 31 empirical equations are reviewed,

and it is observed that most of the equations are unique for certain data gathered from a specific location, while some of the equations perform well. The empirical equations summarized for the prediction of UCS in sandstone vary by input values such as interval transit time, aka P-wave velocity (Fjær et al. 1992; McNally 1987; Moos et al. 1999), porosity (Vernik et al. 1993), and modulus (Bradford et al. 1998). Also, the empirical equations to predict UCS values in shales from porosity (Horshud 2001; Lashkaripour and Dusseault 1993), velocity (Horshud 2001; Lal 1999), and modulus (Horshud 2001) are reviewed and listed in the study. An example of well-performing equations is a relation between strength and porosity for sandstone and shale. The study also emphasized that the velocity readings were from dry rock samples, which causes a lower estimated value of UCS because of the inaccuracy introduced by the difference between dynamic and static moduli. However, it is also noted that the empirical equations derived by using the laboratory results are sufficient to estimate the lower boundary for UCS.

The study conducted by Yurdakul et al. (2011) compared the predictive models for UCS of carbonate rocks from Schmidt Hardness. The Schmidt hammer that was initially designed to measure the strength of concrete can also be used to predict rock strength. The test considers the distance traveled by the energy transferred by the spring, and it measures the Schmidt hardness value based on the percentage of initial extension of the spring. The study compares the prediction results from the first-degree polynomial simple regression model and the artificial neural network (ANN) based model. The data set for this study was collected from 37 different natural stones collected from 19 different natural processing plants from different cities in Turkey. The comparison between models was made considering Variance account for (VAF), coefficient of determination (R^2), and root mean square error (RMSE). A lower value of RMSE indicates a more accurate prediction, while a lower value of VAF indicates a less accurate prediction in the model. Also, R^2 value should be closer to 1 if the model perfectly fits the available data. The obtained VAF, RMSE, R^2 indicators for the simple regression model are 12.45, 46.51, and 0.39, while 95.84, 7.92, and 0.96 for the ANN-based model. The results showed that the ANN-based model performs significantly better than the simple regression model, and an updated model can be developed to predict UCS values in sedimentary rocks.

The performance of various ML methods that can predict UCS is compared in the research completed by Barzegar et al. (2016). The focus of the study is described as to evaluate the performance

of Adaptive neuro-fuzzy inference system (ANFIS), Mamdani fuzzy logic (MFL), Multi-layer perception (MLP), Sugeno fuzzy logic (SFL), and support vector machine (SVM) for the prediction of UCS of rocks in the Azarshahr area in northwest Iran. The fuzzy logic is described as an approach to computational methods that consider the degree of truth rather than absolute truth, and this allows the fuzzy logic to provide arrays of possible true values. The multi-layer perception is a common ANN approach for the prediction models that include layers to process data and learn from it. The adaptive neuro-fuzzy interference system is summarized as a feed-forward neural network function to check for the best fuzzy decision rule. The support vector machine is a soft computing learning algorithm mainly used for classification, pattern recognition, regression analysis, and prediction. The data set for the study include P-wave velocity, porosity, Schmidt rebound hardness, and UCS measured in the laboratory from 85 core samples. For the models, the data set is divided into two subsets: training (80% of data) and testing (20% of data). The performance of the models was assessed based on root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The results indicated that SVM model outperformed the other models with the lowest RMSE (2.14 MPa), MAE (1.351 MPa), and the highest R^2 (0.9516).

Negara et al. (2017) introduced elemental spectroscopy to consider grain size effects on UCS. The support-vector regression (SVR) is utilized to predict UCS. In this study, laboratory testing is the primary method to collect UCS data. The X-ray fluorescence (XRF) analysis is used for elemental spectroscopy. For the models, the data set was collected from the measurements of 35 core samples. The data gathered from seven of these core samples counted as outliers, and only 28 of them were used for the data set. The SVR is a supervised learning method that utilizes the necessary algorithms to analyze and recognize patterns. The quantitative measures to evaluate the performance of the model were the coefficient of determination and the mean absolute percentage error. The results indicated that the model built with SVR could predict UCS with a small error even though a small number of samples were used to train the model. Also, an influence of elemental spectroscopy on UCS prediction is observed. This influence is described as the effect of grain density on rock strength.

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical method that can effectively reduce data dimensionality while preserving the variation within the data set. By preserving the variation, the PCA allows Machine Learning (ML) algorithms to be trained with the same or similar patterns in the data set, which is essential for building a robust ML model. PCA is defined by Gupta et al. (2016) as a dimensionality reduction technique that uses an orthogonal transformation to convert a set of observations of possibly correlated or dependent variables into a set of linearly uncorrelated variables, which are called Principal Components.

Kong et al. (2017) reported that PCA was first published by Pearson (1901) and developed by Hotelling (1933), and the modern applications and concepts are formed by Jolliffe (2002). PCA was used to conduct studies related to history matching, seismic interpretation, pattern recognition, reservoir uncertainty evaluation, data compression, image processing, and high-resolution spectral analysis (Iferobia et al. 2020).

Kong et al. (2017) explained feature extraction as a process of extraction measurements that are invariant and insensitive to the variations within each subspace of the data batch. The feature extraction is an essential step of the task of pattern recognition and the compression of data because both tasks require the smallest possible distortion on data while reducing the number of components. Also, feature extraction is a data processing technique that outlines a high-dimensional space to a low-dimensional space with minimal information loss. Principal component analysis (PCA) is one of the widely known feature extraction methods, while independent component analysis (ICA) and minor component analysis (MCA) are variants of the PCA. ICA is usually applied for blind signal separation, and MCA is commonly used for total least square problems (Kong et al. 2017). The scope of this study will be limited to the PCA, and the analysis will be conducted by using Python (Van Rossum and Drake 2009). However, comprehensive information regarding PCA is provided in Chapter 3 to clarify the concepts.

PCA seems a complicated and time-consuming method once described in mathematical terms, but with the increase in computational power in the last decade, now it is possible to apply PCA to a million data points in less than a minute. With that, the applications of PCA started to become more and more common over the last decade in the industry. The recent applications of PCA include

the spectral decomposition of seismic data, noise reduction of gamma-ray spectrometry maps, predicting possible casing failures, identifying the possible correlations between elemental data, estimation of dominant water channel development in oil wells, and estimation of geomechanical properties in unconventional reservoirs.

Guo et al. (2009) utilized PCA to conduct a spectral decomposition of seismic data technique recently introduced to use as an interpretation tool that can help identify hydrocarbons. The merit of this interpretation technique is to develop an adequate form of data representation and reduction because, typically, the interpreter might generate 80 or more spectral amplitude and phase components. In the study conducted by Guo et al. (2009), 86 spectral components ranging from 5 Hz through 90 Hz were generated using an interpreter, and PCA was utilized to reduce the number of spectral components. It is observed that only three principal components in the total of 86 components were able to capture 83% of the variation in the seismic data. The results indicated that flow channel delineation could be mapped using RGB (Red, Green, and Blue) colors stack for the three largest principal components.

de lima and Marfurt (2018) conducted a study with a similar motivation as Guo et al. (2009). In this study, PCA was used to reduce the noise of the gamma-ray spectrometry maps and reduce the number of components in the data set from four to three. Initially, the gamma-ray spectrometer data consists of TC, K, eTH, and eU. The map displayed after implementing PCA and K-means clustering on PC1 and PC2 indicated a better correlation with the traditional geological map compared to the map created by only clustering of TC, K, eTH, and eU without PCA.

Song and Zhou (2019) conducted a study to predict possible casing failures using PCA and gradient boosting decision tree algorithms (GBDT). The gradient boosting decision tree algorithm is a machine learning algorithm that utilizes decision trees and combines the output of weak and strong decision trees, aka boosting, to create a robust learning algorithm. This study applies the proposed method to the data set obtained from an oil field in mid-east China. The data set was created based on the parameters affecting the casing failure. Some of these parameters were the outside diameter of the casing, the thickness of perforation, and casing wall thickness. The PCA is used to reduce dataset dimensionality, while GBDT is utilized to develop the machine learning classification model. The results indicated that using PCA with GBDT increased prediction accuracy on casing failure compared to classic methods (decision tree, Naïve Bayes, Logistic Regression, multilayer

perceptron classifier). Also, it is stated that the algorithm created by using PCA and GBDT can successfully predict a timeline for preventive maintenance on offset wells (Song and Zhou 2019).

Even though the PCA is commonly used to reduce dimensionality, aka the number of variables in the data set, PCA has other practical applications. The PCA can be utilized to identify a correlation between the components within the data. The study conducted by Elghonimy and Sonnenberg (2021) focuses on observing a correlation between major and trace elements within the elemental data obtained from the Niobrara Formation in the Denver Basin. In the study, elemental data of samples is measured using a handheld XRF analyzer on full core from the Niobrara Formation. The variability of elemental concentrations is analyzed using PCA, and it is compared with the core facies to display the history of deposition and the conditions through the deposition process. The results showed that the application of PCA on the data set created by the integration of XRF measurements and core facies indications made a clear display of these elements in five major categories. Also, it is stated that these identified major categories can act as an intermediary for the different deposited elements to indicate the history of deposition within the Niobrara Formation.

Chen et al. (2019) studied a possible application of PCA as a recognition method for the dominant water channel development in oil-producing wells. The study was conducted using SZ Oilfield's data located in Liaodong of Bohai bay. An evaluation index system was created to build a comprehensive evaluation method to consider every parameter. The parameters grouped in two main categories, dynamic response parameters and the parameters causing the channel to advance. The parameters considered for the evaluation index system are as follows:

- Dynamic response parameters.
 - Dimensionless pressure index,
 - Pressure index,
 - Average water cut,
 - Water absorption profile coefficient,
 - Apparent water injectivity index increase,
 - Water injection intensity increase.
- Parameter causing the channel to advance.

- Total water injection volume/unit thickness,
- Apparent water injection intensity,
- Viscosity of crude oil,
- Effective thickness of reservoir,
- Permeability contrast,
- Average permeability.

In this study, the focus of utilizing PCA is to test the objectivity of the method since the increasing number of parameters induces subjectivity to the recognition algorithm. An evaluation index system is created to analyze the causes of the dominant channel, and based on this system, an artificial learning method that recognizes the dominant channel is developed using PCA. The decision system was created based on the comprehensive evaluation index of the well group. If the calculated comprehensive evaluation index of the well group was higher than the average value, the well group was assumed to be developing a dominant channel. The results showed that the application of PCA to compute comprehensive evaluation index values reduced the subjectivity introduced by a large number of parameters. Also, it is stated that the method can provide technical support for further enhancing oil recovery by recognizing a pattern of dominant channel development in producing wells.

Furthermore, another study conducted by Ifeobia et al. (2020) shows that the significant number of drilling operations conducted in unconventional reservoirs has shown that the prediction of UCS in the shale reservoirs is essential due to its complex and non-linear behavior. However, the models with a single log input parameter(sonic) were insufficient for these formations (Ifeobia et al. 2020). In the study, 21,708 data points of acoustic parameters are used to create a model with a principal component – multivariate regression, and the results indicated that the model could predict UCS values with 99% accuracy.

Previous studies conducted on possible wellbore stability problems induced by a lack of knowledge of geomechanical properties indicated that the geomechanical modeling for drilling and completion operations is open for improvement. Furthermore, the previous works conducted on the correlation between drillability and UCS show that the estimation of UCS is essential to increase

the overall drilling performance. ML algorithms are powerful tools for predicting UCS on formations with complex and non-linear behavior. Recent studies show that data-driven solutions are reliable resources to support the decision-making process for drilling and completion operations. ML can be a powerful tool to estimate geomechanical parameters such as UCS in formations with complex and non-linear behavior. The literature review shows that implementing these methods can vastly increase overall drilling performance and efficiency of completion operations.

2.3 Tree-Based Algorithms

Tree-based algorithms are supervised learning methods that are considered to be the most efficient machine learning method. Tree-based algorithms can be used for both regression and classification problems. Also, they are suitable for non-linear relationships. The tree-based algorithms are simple yet powerful learning methods. The most popular tree-based learning algorithms are decision trees, gradient boosting, and random forest. Tree-based methods create and partition the feature space into a set of subspaces and fit a simple model (or constant) in each space. By using these feature spaces, a decision for each entry is given based on conditions set on each node of every tree.

The decision tree is a supervised learning algorithm with a defined target available. The decision tree is commonly used for classification problems. There are two types of decision trees: categorical variable decision trees and continuous variable decision trees. Categorical decision trees are built to solve classification problems, while continuous variable decision trees are commonly used to solve regression problems. Similarly, for both decision trees, the algorithm splits the data set into two or more homogenous subsets regarding the highest number of splitters or differentiators in input values. The decision tree algorithms are popular as they are easy to understand and useful for data exploration, but their tendency to overfit the data set is the most common difficulty of these methods.

Regression and classification trees are simply decision trees with more nodes (or leaves), splitting the data set into smaller subsets. As it is mentioned before, the main differences between them are the type of input values and objective set while training the algorithm. While training, the splitting process creates a fully grown tree for both cases, and the split process can cause overfitting as the information given to each tree will be similar. The model parameters can be adjusted to

avoid overfitting, and validation techniques like pruning can be applied. The constraints of the tree size are simply the model parameters such as minimum samples for each node, minimum samples for each terminal node, maximum features to consider for split, and the maximum depth of the tree. The pruning essentially prevents the model from being greedy in the decision process. While pruning, the tree is grown to a large depth, and nodes (or leaves) that give negative values as output are removed.

Another advantage of tree-based algorithms is that they are suitable for ensemble learning methods. Ensemble learning is developing a group of predictive models to improve model stability and accuracy, and they are simply a boost for decision tree models. A well-developed model should maintain the balance between bias and variance error. This balance between two errors is known as the trade-off management of bias-variance errors. This trade-off between variance and bias can be optimized in decision tree models by applying ensemble learning methods. The most common ensemble learning methods are bagging, random forest, and boosting. The bagging and random forest are the most common methods applied for classification problems where a group (committee) of trees cast a vote for the prediction. Boosting also involves a committee of trees that cast a vote for the prediction, but unlike random forest, a group of weak learners can evolve, and committee members cast a weighted vote. Bagging also utilizes multiple classifiers and combines them to develop a classifier to reduce the variance of predictions. These ensemble methods simply develop a group of base learners and combine them to establish a strong composite predictor. Even though there are many ensemble learning methods, Random Forest (RF) is decided to be the most suitable algorithm for this study, and it will be used to build a regression model to estimate UCS from drilling parameters.

2.3.1 Random Forest

Random forest is an ensemble learning method that includes multiple classifiers and uses a combination of classifiers. Random forest is technically a modification of the bagging method that utilizes many decorrelated trees and retains the average of predictions as an output. Random Forest (RF) is similar to Boosting in terms of performance, but tuning the hyperparameters to avoid overfitting makes RF the best option for this study.

The algorithm of RF for regression or classification is as follows,

- Assume $z=1$ to Z :
- Then a bootstrap sample A^* of size N is drawn from the training data.
- RF tree T_z is grown around the bootstrapped data by following the steps below for each terminal node of the tree until the predetermined node size is achieved.
 - n number of variables are selected from the p variables.
 - The best variable point among the n is picked.
 - The node is split into two daughter nodes.
- The output of ensemble trees comes as $\{T_z\}_1^Z$

The prediction at a new point x ;

$$\text{Regression : } f_{rf}^Z(x) = \frac{1}{Z} \sum_{z=1}^Z T_z(x) \quad (2.4)$$

Classification : If the class prediction of the z^{th} random forest tree assumed as $C_b(x)$ (2.5)

$$\text{Then } C_{rf}^Z(x) = \text{majority vote } \{C_z(x)\}_1^Z(x) \quad (2.6)$$

The process of bagging trees creates identically distributed variance within the predictions, which means that the bias of a single tree (bootstrap) is the same as the bagged trees. Therefore, the only hope of improving prediction accuracy is to reduce the variance. However, RF adaptively develops the trees to remove bias as the distribution between trees is not identical.

If the variance is explained by the terms of the RF algorithm described above. Then, the average variance of Z independent identically distributed random variables, each with σ^2 variance, is $\frac{1}{Z}\sigma^2$. If it is assumed that the variables are identically distributed but not necessarily independent from each other, with a positive correlation ρ , the average variance is

$$\rho\sigma^2 + \frac{1-\rho}{Z}\sigma^2 \quad (2.7)$$

The second term of the average variance disappears with the increasing number of random variables (Z). Hence the averaging greatly helps to reduce the variance. The focus of RF is to

reduce the variance of bagged trees by reducing the correlation between them without increasing variance by a high margin.

RF is one of the most common ML methods that help solving to both regression and classification problems. Also, the studies reviewed indicated that the common problems of the oil and gas industry, specifically the drilling industry, can take the opportunity to create models with RF that can predict essential parameters or indicate the drilling incidents.

Hegde et al. (2015) conducted a study to oversee the possible applications of statistical learning techniques to predict the Rate of Penetration (ROP) values. The statistical learning methods were trees, bagged trees, and Random Forest (RF), and the models are evaluated based on their performance by comparing Root Mean Square Error (RMSE) values. Also, this study introduces a predictive model called Wider Windows Statistical Learning Model (WWSLM), which considers many input parameters such as WOB, RPM, and depth to compensate for the effect of high lithology variation on drilling parameters by utilizing trees and random forest. Another advantage of the model built by Hegde et al. (2015) is that the blind test subset is split from the complete data set and ensured that the model never learns from it, which reduces the probability of overfitting. Only surface drilling parameters are included in the training, validation, and test sets as input data. The tree method is described as a technique used for classification and regression purposes. Bagging is summarized as jointly using bootstrapping and decision trees to compensate for the high lithology variation. Bootstrapping is a method to reduce variation by repeatedly sampling from the same data set that yields multiple training sets. Random forest is described as a method that utilizes bootstrapping to increase the number of training subsets and combines it with decision trees. The main difference between RF and bagging is that RF uses a random subset of predictors to build trees, which eventually reduces the variance of Statistical Learning Model (SLM). The results indicated that RF provides the most accurate ROP predictions with RMSE of 7.4, which is three times lower than RMSE values of other models. It should be noted that the ROP values of the training data set ranging from 20 to 120 ft/hour.

The optimization of drilling parameters by using Random Forest (RF) regression algorithm is studied by Nasir and Rickabaugh (2018). The study aimed to optimize drilling parameters to increase bit life by reducing the wear and tear while maximizing ROP and minimizing Mechanical Specific Energy (MSE). In the study, drilling parameters from the wells within 20 miles radius are

used as a training data set. Also, it should be noted that the drilling parameters were collected while using the same motor and bit for the entire vertical interval of these wells. The key drilling parameters investigated were surface RPM, mudflow rate, WOB, and formation type. A total of six different formations are encountered while drilling these wells. The models were trained to estimate ROP using Linear regression, Support Vector Regression (SVR), Random Forest, and Boosted Tree. The data is split into two categories, the training set (80% of the data set) and the validation set (20%). Performance indicators for the model were root mean square error (RMSE), mean absolute error (MAE) and mean average error percentage (MAEP). The results showed that RF could estimate the ROP values with 12% error after tuning the hyperparameters, while the other models had a higher percentage of errors. Also, the author stated that the model could be possibly improved by introducing key variables impacting drilling performance, such as compressive strength and gamma-ray response.

AlSaihati et al. (2021) studied the possible application of Random Forest (RF) to predict the anomalies in Torque and Drag (T&D) values to indicate a possible drilling incident such as a stuck pipe in advance. While building RF model, a pipe stuck case that took place while drilling a 5-7/8-inch horizontal section of 15,060 ft depth well is considered, and the model was built to indicate possible problems in this interval. It is assumed that the stuck pipe occurred while drilling the reservoir contact zone (5000 ft) because of high T&D, insufficient transfer of weight-on-bit (WOB), and poor hole cleaning. The pipe stuck occurred at a measured depth of 14,935-ft while tripping out after circulating the hole clean. The drilling parameters used as a variable to train RF model include hook load, flow rate, rotation speed, rate of penetration, standpipe pressure, and torque readings at the surface, and the data covers the timeline starting from the beginning of the horizontal section to one day prior to the stuck pipe incident. The number of data points used to train the RF algorithm is 7,186, 80% of the total data, and 1,797 data points used to test the algorithm, which refers to 20% of the total data. The MSE value of 0.06 and R value of 0.97 indicated that the RF algorithm could predict the anomalies accurately. Also, it is stated that the model detected anomalies for nine hours consecutively prior to the pipe stuck incident.

Overall, the reviewed studies indicated that the performance of RF machine learning algorithm could be a robust tool, and its implementation as a solution for common problems is promising. Also, the studies indicated that RF is one of the most common algorithms to build efficient data-

driven solutions for the oil and gas industry. RF algorithm can potentially be implemented to estimate geomechanical and operational parameters for drilling automation systems as long as the collected data is accurate.

CHAPTER 3

BACKGROUND INFORMATION ABOUT DATASET AND PCA

Data collection and processing are essential for every project studying possible data-driven solutions for the common problems. In this research, the drilling data used to train a regression model to estimate UCS is collected for another research project with similar intents (Joshi 2021). In this chapter, the background information of the data set is provided. Also, the unique nature of this data set is discussed in detail. In addition, the fundamental principles of PCA are discussed in this chapter. Then, the implementation of PCA to indicate feature importance and explained variance by principal components are given in detail with the results.

3.1 Background Information about the Dataset

As mentioned before, the data set needs to meet certain criteria to be used to train a machine learning algorithm. In other words, the accuracy, consistency, uniqueness of data, and completeness and validity of the data set should meet certain criteria. The possible problems in a data set can be duplicated, outlier, missing data points, and inaccurate data readings. Initially, the data set used in this research was collected by Dr. Joshi, and it was collected while drilling through samples prepared in laboratory conditions. While the analog samples were prepared to mimic field conditions while drilling, cryogenic samples were prepared to mimic various extraterrestrial conditions in Lunar. Joshi (2021) collected the drilling data using a laboratory experiment setup. The properties of these analog and cryogenic samples are given in Table 3.1 and Table 3.2, respectively. The initial plan was to use the complete raw data set collected from each of these samples, but unfortunately, a drilling data collected only from analog samples could be utilized to train the RF regression model for this study. A discussion regarding the uniqueness and challenges of the data set is provided later in this chapter. The final form of these cellular concrete samples, aka analog samples, before and after drilling while curing and demolding is presented in Figure 3.1.

The setup is built on a frame with a 3-phase AC motor to transfer power to the masonry bit. The movement of the masonry bit is supplied by the stepper motor and guide rails. The experimental setup was controlled by a variable frequency drive and stepper motor. The picture of

Table 3.1: Properties of Cryogenic Samples. Modified from (Joshi 2021)

Type of Sample	Water-ice Content (%)	Number of Samples
Low - porosity aqueous regolith simulant	3%	2
	6%	2
	9%	2
	Layered 3% - 6% - 9%	1
	Layered 9% - 6% - 3%	1
High - porosity aqueous regolith simulant	3%	2
	6%	2
	9%	2
Unfused granular icy regolith simulant	3%	1
	6%	1
	9%	1
Fused granular icy regolith simulant	3%	1
	6%	1
	9%	1

Table 3.2: Properties of Cellular Samples. Modified from (Joshi 2021)

Sample	Density (gr/cc)	Dimensions (cm^3)	Strength(MPa)	Porosity(%)
Sample A	1.048	44*44*11	8.69	37%
Sample B	0.850	44*44*15	5.89	49%
Sample C	0.705	44*44*15	3.93	58%
Sample D	0.497	44*44*15	1.50	69%
Sample E	0.329	44*44*14	0.53	78%

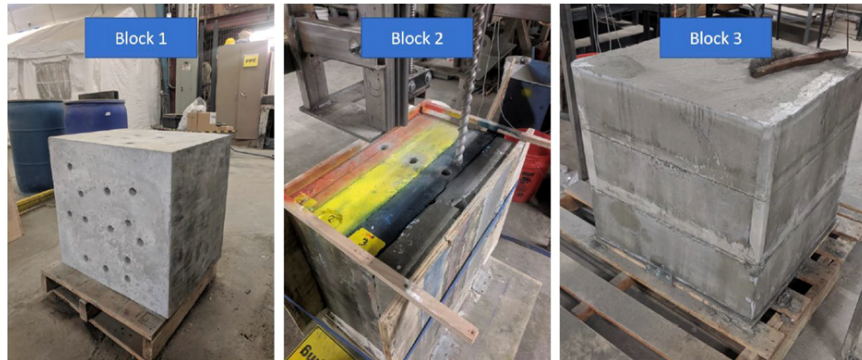


Figure 3.1: Three Concrete Blocks while curing and demolding.(Joshi 2021)©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi

the experimental setup is presented in Figure 3.2.

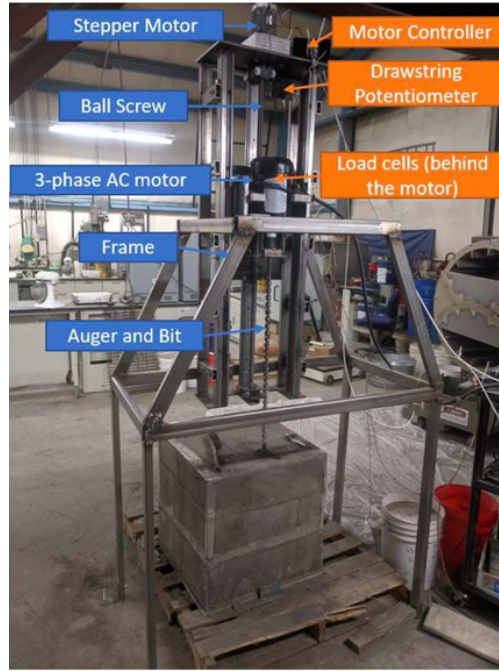


Figure 3.2: Experimental Setup (Joshi 2021) ©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi

The cDAQ data acquisition system and LabVIEW VI are used to measure and record the drilling data. Four parameters are measured using the Data Acquisition System (DAQ): depth, RPM, axial forces, and time. The rest of the drilling parameters: torque, WOB, ROP, Normalized Field Penetration Index (N-FPI), and MSE, are derived from the four measured parameters (Joshi 2021). The following equations are used to calculate these derived parameters.

3.1.1 Drilling Depth

The drilling depth calculations were based on the drilling head height or position sensor (Joshi 2021). The head height at the top of the sample is considered as initial point, and drilling depth at any point is calculated as:

$$Drilling\ Depth_i = Drilling\ head\ height_i - Drilling\ head\ height_{i-1} \quad (3.1)$$

3.1.2 Weight on Bit (WOB)

Weight on bit calculations was based on the subtraction of axial force measurement on air for each bit from axial force measurements while drilling. The reference axial force of each bit on-air is measured and hardcoded to the model to automate the WOB calculation process (Joshi 2021). The WOB at any point is calculated as:

$$WOB_i = Total\ axial\ force_i - Total\ axial\ force_{air} \quad (3.2)$$

3.1.3 Rate of Penetration (ROP)

The rate of penetration is calculated by the difference in measured depth at each time interval. The ROP can be calculated by the following equation.

$$ROP_i = \frac{Drilling\ Depth_i - Drilling\ Depth_{i-1}}{time_i - time_{i-1}} \quad (3.3)$$

3.1.4 Normalized Field Penetration Index (N-FPI)

The field penetration index was originally defined to evaluate the energy required to overcome the rock strength for a tunnel boring machine (Tarkoy and Marconi 1991), (Hassanpour et al. 2011). By the original definition, the FPI is calculated as:

$$FPI = \left(\frac{\frac{kN}{cutter}}{\frac{mm}{rev}} \right) = \frac{F_n}{P} \quad (3.4)$$

Where, F_n cutter load (Cutter Load (kN)/Number of Cutters) and P is the penetration of cutter per revolution (ROP/RPM).

The cutter load (F_n) can be replaced by the normalized drilling force (WOB) to calculate Normalized Field Penetration Index (N-FPI) for the drilling systems. The cutter load F_n for drilling systems can be calculated as:

$$F_n = \left(\frac{N}{mm^2} \right) = \frac{WOB}{\left(\frac{\pi}{4} \right) d^2} \quad (3.5)$$

The final unit of N-FPI will be $\left(\frac{N}{mm^2} \right) / \left(\frac{mm}{rev} \right)$

3.1.5 Mechanical Specific Energy (MSE)

Teale (1965) defined the Mechanical Specific Energy (MSE) as the amount of energy required to excavate a unit volume of rock. MSE can be calculated as:

$$MSE = \left(\frac{MJ}{m^3}\right) = \frac{WOB(N)}{\left(\frac{\pi}{4}\right)d^2(mm^2)} + \frac{10^3 Torque(N.m) \times RPM}{\left(\frac{\pi}{4}\right)d^2(mm^2) \times ROP(mm/min)} \quad (3.6)$$

3.2 Uniqueness with Dataset and Challenges

Every data set collected includes noise and outliers that need to be removed before using for any machine learning or deep learning application. The process of filtering these noise and outliers from raw data before labeling is called processing. The challenges introduced by raw data can vary, and its uniqueness is the key to building a robust machine learning model. Likewise, the data set used in this study needed processing before being used as training data for machine learning models, but the challenges were not limited to only noise. Initially, torque values were planned to be measured using a bridge-based shaft-to-shaft torque sensor placed between shaft and bit, but torque readings were incredibly noisy due to electromagnetic interference, mechanical noise, and ambient noise (Joshi 2021). The electromagnetic interference caused by the three-phase AC motor is located extremely close to the torque sensor. The vibrations in the bit while drilling result in mechanical noise in torque readings, while ambient noise from The Earth Mechanics Institute (EMI) at the Colorado School of Mines campus was detected by torque sensors. Filtering the noise introduced by the environment and experimental setup was removed in a sampled data, but this process required a vast amount of time as signal smoothing, outlier removal, and significant signal filtering needed to be applied to the complete data set. After consideration, Joshi (2021) decided to build a regression model to predict torque values from other drilling parameters and Variable Frequency Drive (VFD) outputs, as the filtering process was computationally expensive and required significant signal conditioning. To clarify how torque data is calculated, an architecture of the algorithm called “The Lunar Material Characterization while Drilling Algorithm” is presented in Figure 3.3.

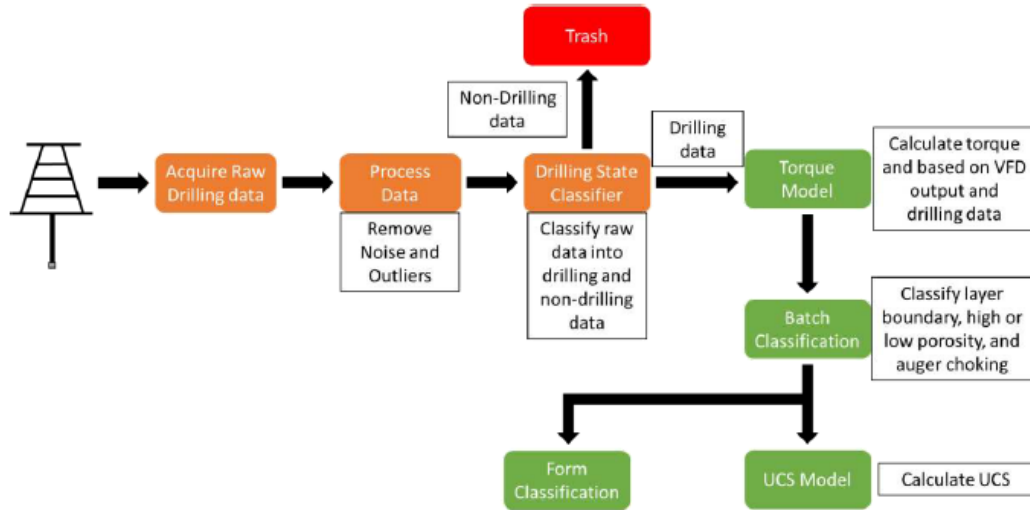


Figure 3.3: The architecture of algorithm built by Joshi (2021) ©2021 by Deep R. Joshi., reprinted with permission from Deep R. Joshi

As it is described in Figure 3.3, a regression model used to calculate Torque values from VFD output and other drilling parameters right after raw data is classified as drilling and non-drilling data. Later, Torque values were assumed to be similar to torque sensor readings. Another challenge introduced by this dataset is the lack of variation within UCS.

Even though the complete data set contains approximately 55+ observations on eight different variables, only seven different UCS values are present. Initially, this was assumed to be a unique part of the data set. Later, it was understood that the lack of variability in target values while building RF regression model was causing significant overfitting and leading to high prediction accuracy while estimating UCS. A histogram graph of UCS within the complete data set is given in Figure 3.4. This issue and its impacts on this study are discussed in detail in Chapter 4. Initially, in the scope of this study, these predicted torque values were assumed as a unique part of the data set. However, the prediction of torque values possibly reduced the variation within the data set and impacted this study significantly with the contribution of other data set challenges. This combination led to changes in the path of this study.

3.3 Principal Component Analysis

Brief information about Principal Component Analysis (PCA) is given in Chapter 2. PCA is defined as a dimensionality reduction method that utilizes orthogonal transformation to convert

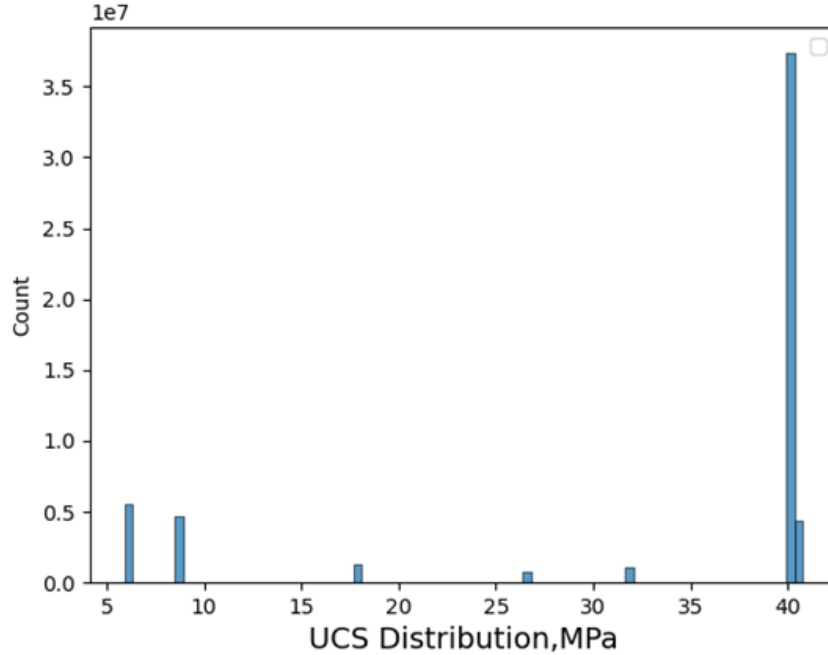


Figure 3.4: Histogram of UCS - Complete Dataset

a set of observations into a set of linearly uncorrelated variables (Gupta et al. 2016). As is described comprehensively in Chapter 2, PCA is commonly used to build data-driven solutions for the various problem caused by dimensionality. This section explains the fundamental principles and implementation of PCA to clarify the concept. In this study, the scope of PCA will be limited to feature importance indication and explained variance analysis in the data set.

3.3.1 The Concept of PCA

Kong et al. (2017) stated that “the principal components (PC) are the directions in which the data have the largest variances and capture most of the information contents of data.”. The PCs are correlated with the eigenvectors inherent in the largest eigenvalues of the autocorrelation matrix of the data vectors. The expression of data vectors regarding the PCs is named PCA, while expressing the vectors regarding the MCs is named MCA.

PCA or MCA is usually one-dimensional, but the actual applications have also been found to be multiple dimensional. The principal (or minor) components are referred to as the eigenvectors affiliated with r largest (or smallest) eigenvalues of the autocorrelation matrix of the data vectors, while r is the number of the principal (minor) components. The subspace covered by PC is called

principal subspace (PS) or signal subspace, while the subspace covered by MC is called minor subspace (MC) or noise subspace. The PCs and MCs result from converging matrix differential equations derived from the symmetric matrix's principal and minor component analyzers (Kong et al. 2017). Since the PC is the direction correlated with the eigenvector affiliated with the largest eigenvalue of the autocorrelation matrix of the data vectors and the MC is the direction correlated with the smallest eigenvalue of the autocorrelation matrix of the data vectors, batch eigenvalue decomposition (ED) of the sample correlation matrix or singular value decomposition (SVD) of the data matrix can be used for PCA and MCA.

The concept of PCA can be described with math equations. The goal of PCA is to reduce the dimensionality of the matrix $X \in R^{N \times D}$ to matrix $Z^{N \times M}$ by projecting X into lower dimension space M where $D \gg M$.

$$X = [x_1 x_2 x_3 \dots x_N]^T \quad X \in R^{N \times D} \quad (3.7)$$

$$Z = [z_1 z_2 z_3 \dots z_N]^T \quad Z \in R^{N \times M} \quad (3.8)$$

By using transformation matrix U, the dimensionality of matrix X should be reduced from D to M.

$$Z = U * X \quad U \in R^{D \times M} \text{ where;} \quad (3.9)$$

$$\text{The covariance matrix of } S = S_Z = \frac{1}{N} Z^T * Z \quad S_Z \in R^{M \times M} \quad (3.10)$$

Optimization should be done to maximize the covariance matrix S, and an upper boundary condition should be added to avoid an infinite number of results.

$$\max_u S_Z \quad (3.11)$$

$$\max_u \frac{1}{N} Z^T * Z \quad \text{where;} Z = UX \quad (3.12)$$

$$\max_u \frac{1}{N} (XU)^T (XU) \quad (3.13)$$

$$\max_u \frac{1}{N} U^T X^T XU \quad \text{where;} X^T X = S_x \quad (3.14)$$

$$\max_u \frac{1}{N} U^T S_x U \quad (3.15)$$

The boundary condition is that every vector in this matrix has a unit magnitude.

$$U^T U = 1 \quad (3.16)$$

Now that there is an optimization problem with equality constraints, it can be solved using LaGrange multipliers.

The general application of LaGrange multipliers is as follows.

$$L(x, \{\lambda_i\}) = f(x) + \sum_{i=1}^n \lambda_i g_i(x) \quad \text{where; } \lambda_i \geq 0 \quad (3.17)$$

$$\frac{\delta L}{\delta x} = 0 \quad (3.18)$$

$$x = h(\{\lambda_i\}) \quad (3.19)$$

For the target equation of PCA.

$$L(U, \lambda) = U^T S_x U + \lambda(1 - U^T U) \quad (3.20)$$

$$\frac{\delta L}{\delta x} = 0 \quad (3.21)$$

$$S_x U = \lambda U \quad (3.22)$$

$$S_i U_i = \lambda_i U_i \text{ (Eigenvector Equation)} \quad (3.23)$$

As mentioned before, the methods of solving the eigenvector equation are Eigenvalue decomposition (ED), aka matrix diagonalization, and Singular value decomposition (SVD). The mathematical explanation of these methods is summarized below.

3.3.1.1 Eigenvalue Decomposition (Matrix Diagonalization)

The concept of ED involves the decomposition of a square matrix into three different product matrices. For such matrix $S \in R^{N \times N}$

$$S = P D P^{-1} \quad S, P, D \in R^{N \times N} \quad \text{where;} \quad (3.24)$$

The matrix of eigenvectors P is,

$$D = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_N \end{bmatrix} \quad (3.25)$$

$$\text{where; } SU = \lambda U \quad (3.26)$$

The pairings of eigenvalues and eigenvectors are sorted out in descending order to select the values representing the retained explained variance most,

$$(\lambda_1, U_1), (\lambda_2, U_2), \dots, (\lambda_D, U_D) \quad \text{and} \quad \lambda_1 > \lambda_2 > \dots > \lambda_D \quad (3.27)$$

The reason to find these pairs of values is to determine the variance of data represented by eigenvalues. Furthermore, the variance of data represented by different eigenvalues can be calculated by:

$$\frac{\text{Retained Variance}}{\text{Total Variance}} = \% \text{information represented by data} = \frac{\sum_{i=1}^D \lambda_i}{\sum_{i=1}^N \lambda_i} \quad (3.28)$$

3.3.1.2 Singular Value Decomposition

The history of SVD dates back to the 1870s. Beltrami (1873) published a study on SVD, Jordan published his reasoning about SVD (1874). Later, the concept of SVD covered complex square matrix with the study conducted by Autonne (1902), and the theorem even extended further to cover the general rectangle matrix by Eckart and Young (1939) (Kong et al. 2017). Currently, the SVD theorem for rectangle matrix is commonly named Eckart-Young Theorem.

Assume a design matrix $X \in \mathbb{R}^{N \times D}$; this design matrix can be decomposed into three different matrices such that $U \in \mathbb{R}^{N \times N}$, $S \in \mathbb{R}^{N \times D}$ and $V \in \mathbb{R}^{D \times D}$. The matrix S is assumed to be a diagonal matrix that contains the singular values of the design matrix.

$$X = USV^T \quad (3.29)$$

The U and V matrices have orthonormal columns, so

$$I = UU^T \quad \text{and} \quad I = VV^T \quad (3.30)$$

$$X^T X = VS^T U^T U S V^T \quad (3.31)$$

$$X^T X = VS^T S V^T \quad \text{where; } D = S^T S \quad (3.32)$$

$$(X^T X)V = VD V^T V \quad (3.33)$$

$$(X^T X)V = VD \quad \text{where; } \quad (3.34)$$

$$Cov(X) = \frac{1}{N} \sum_{n=1}^N X X^T \quad (3.35)$$

The equation above shows that $X^T X$ is the scaled version of the covariance matrix of design matrix X . The matrix V contains the eigenvectors of the design matrix, and the matrix D contains the eigenvalues of the design matrix. By applying the same process as eigenvalue decomposition, the information represented by each eigenvalue can be calculated.

3.3.2 The Results of PCA

The fundamental concept of principal component analysis is explained with mathematical expressions in the previous section. As mentioned before, the main objective of PCA is to measure the variance within the data set and provide insight into how the variance contribution changes with each principal component. The measurement of variance contribution by each principal component is called explained variance analysis. Also, PCA can be used to indicate the variance contribution of each variable by simply measuring the number of data points transferred to each principal component. This process is called Feature Importance Analysis, and it can be used to indicate the influence of each variable on principal components. In this study, both explained variance analysis and feature importance analysis will be applied to the complete data set to indicate explained variance by each PC and how these PCs are influenced by each variable in the data set. Implementation of PCA helped to gain insight into how the variance within the data set changes with each variable, and the insight eventually led to changes in this study. The PCA will be applied only to the training data set as the objective is to study the variance contribution of each variable. These variables are MSE, ROP, Torque, N-FPI, WOB, RPM, and Depth.

3.3.2.1 Explained Variance Analysis

Explained variance can be defined as the variance retained by each eigenvalue and eigenvector described by the covariance matrix of principal components. . These retained values by each principal component is calculated by using scikit-library (Pedregosa et al. 2011). The explained variance of each component is indicated in Figure 3.5. As can be observed in Figure 3.5, PC1 reflects the highest variance within the dataset as the eigenvalues and eigenvectors are sorted out after applying eigenvalue decomposition. The observation from this graph is the uniform decrease

in explained variance by each principal component. An exponential decrease should be observed after the first two principal components, as they will carry the highest amount of variance within the data set. This exponential decrease phenomenon can be observed in explained variance in iris dataset in Figure 3.6 (Pedregosa et al. 2011).

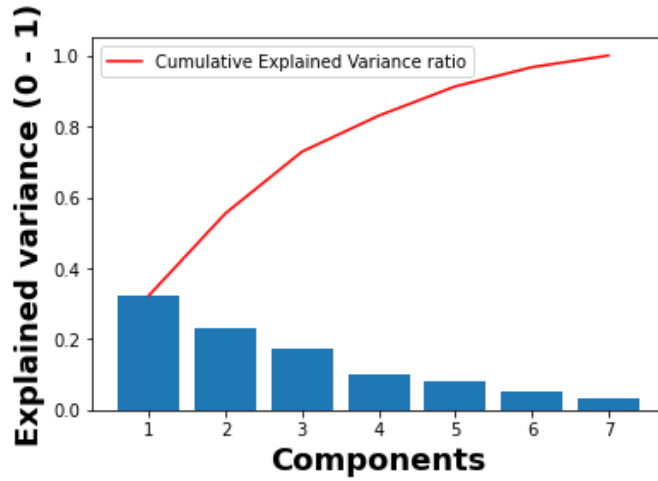


Figure 3.5: Explained Variance within the Training Dataset

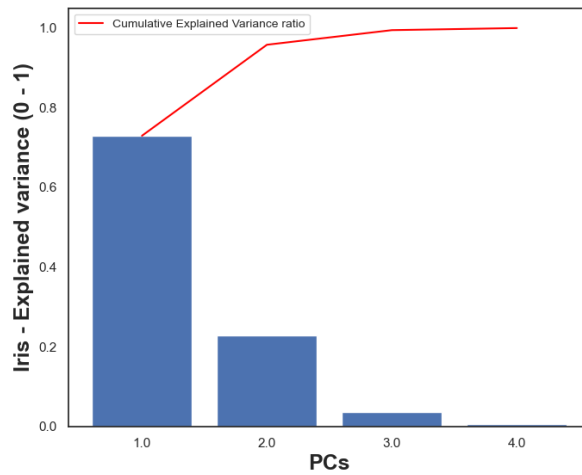


Figure 3.6: Explained Variance within Iris Dataset (Pedregosa et al. 2011)

3.3.2.2 Feature Importance Analysis

A feature matrix indicates the importance of each feature that is reflected by the magnitude of corresponding values in eigenvectors. The numbers are shown in the feature matrix range from -1 to 1, indicating the influence of features on each principal component. The higher number indicates a more significant influence. The feature matrix of the training data set is presented in Figure 3.7. The coefficient values reflect the influence of each variable on principal components. The calculated coefficient value can be negative or positive since these values indicate a distance, aka magnitude, from orthogonal lines set for each principal component. The observation from the feature matrix is a significant influence of variables on PC7. Commonly, the degree of influence decreases with each principal component added, but similar to the anomaly observed from explained variance percentages, the influence of variables on principal components is not decreasing. Especially, the influence of MSE and Torque are high on PC7, which indicates that valuable information from the original data set will be lost if less than seven PCs are used to train the algorithm.

The explained variance and feature importance analysis indicated that a possible implementation of PCA to reduce dimensionality will result in the loss of valuable information. This observation is simple yet essential as it impacted this study and led to changes. A detailed discussion on the possible loss of information is given in Chapter 4.

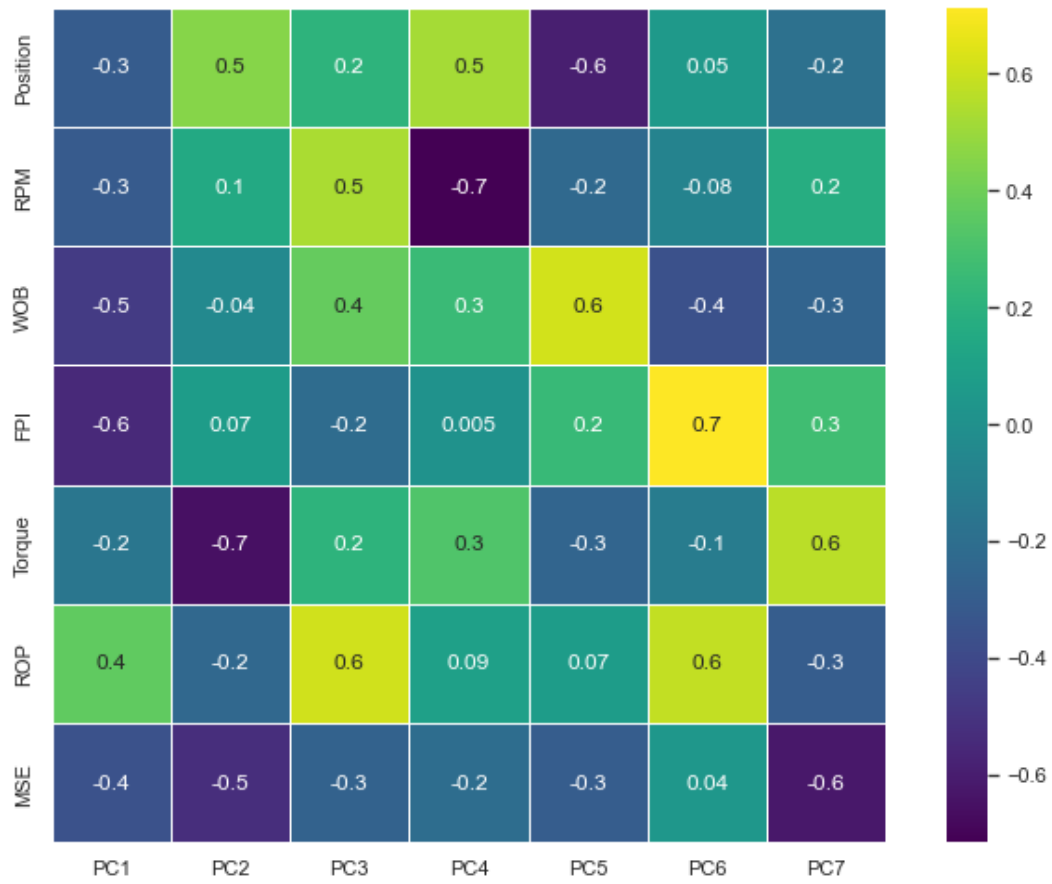


Figure 3.7: Feature Matrix indicating influence of each variable to PCs

CHAPTER 4

MODELING AND TESTING

Once negative values are removed and the data set re-indexed successfully, the drilling parameters are fed into the Random Forest regression algorithm as a training data set. The objective is to develop a regression model that estimates unconfined compressive strength from drilling parameters. This chapter covers machine learning methods and includes model performance evaluation methods while building a regression model. Also, the final architecture of the Random Forest regression algorithm is explained.

4.1 Machine Learning Methods

Machine learning is stated as the science of programming computers so they can learn from data (Geron 2019). Machine learning can be a valuable asset for solving extremely complicated problems that can not be explained using traditional physics-based models. Due to drilling operations' complex and dynamic nature, machine learning methods to predict geomechanical and operational parameters perform better than traditional models that commonly consider several assumptions. The machine learning methods can be grouped into two main categories, supervised learning and unsupervised learning. The main distinction between supervised and unsupervised learning methods is the data set used to train models. Supervised learning methods use labeled data sets for classification and regression problems, whereas unsupervised learning methods use unlabeled data sets for clustering, association, and dimensionality reduction problems.

This study uses a supervised learning method called Random Forest regression algorithm to estimate UCS instantaneously from drilling parameters. The algorithm built includes two main sections. The first section consists of a main Random Forest regression algorithm and hyper-parameter tuning tools for RF model to ensure the best fit while avoiding overfitting. These hyper-parameter tuning tools are called Randomized-Search and Grid-Search, which have a built-in K-fold cross-validation function. The K-fold cross-validation function is essential while tuning hyper-parameters to ensure the best bias-variance trade-off. The second section includes common model performance evaluation techniques such as root-mean-squared-error, and mean-absolute-

error. In this chapter, the most common regression models and performance evaluation techniques are evaluated, the algorithm's final architecture is provided, and the potential implementation of the model for drilling operations is discussed.

4.1.1 Regression Models

Regression problems can be defined as problems that require a quantitative approach to solve. A quantitative relation between one or more dependent or independent features can be measured using regression models. Regression models can be used to solve the simple or complex relations between one or multiple features. The regression model complexity will increase with the increasing complexity of relations between the features and target values. This relation can be as simple as linear, which can be solved with a first-degree polynomial equation, or incredibly complex, requiring more sophisticated methods such as decision trees. More information regarding these sophisticated methods is given as four different regression models are evaluated in this section.

4.1.1.1 Linear Regression

Linear regression is the most common and simplest regression method that can be defined as the simplest supervised machine learning algorithm for regression problems. The lack of complexity of linear regression can be an advantage as long as the relation between dependent and independent features can be explained with a first-degree polynomial equation. Linear regression is still robust and practical approach to building simple quantitative models. The main linear regression types can be divided into simple and multiple linear regression. Simple linear regression is a powerful regression method when a parameter is estimated using only one predictor or feature, aka a single independent variable. Even though simple and multiple linear regression methods follow similar techniques to estimate features, a multiple linear regression provides the complexity needed to estimate more than one feature. The multiple linear regression model is required to estimate UCS from drilling parameters as more than one feature is present in the data set. A mathematical description of the multiple regression method is given below in Equation 4.1.

$$Y = \beta_0 + (\beta_1 X_1) + (\beta_2 X_2) + \beta_3(X_1 * X_2) + \beta_4(X_1 * X_2 * X_3) + \dots + \epsilon \quad (4.1)$$

where:

Y: Output Parameter or Predictor

X_i : Input Parameters or Features

β_i : Regression Coefficients

ϵ : Error

Here m_i can be defined as average impact of X_i or $X_i * X_{i+1}$ on Y.

The goal of implementing a multiple regression model is to estimate the regression coefficients (β_i) for each input parameter or feature while satisfying the requirement of having the minimum sum of square residuals. This simple approach performs well while building a robust multiple regression model, but it is essential to measure the feature importance to recognize a relation between features and predictors. By recognizing the relation between features and predictor, features with a small impact on predictor can be eliminated. The most common method to estimate feature importance is the null hypothesis. The null hypothesis is essentially replacing each regression coefficient with integer zero to identify the importance of each feature. Mathematically, the null hypothesis for a model with n features can be described as (James et al. 2013)

$$H_o = \beta_0 = \beta_1 = \beta_2 = \dots = \beta_n = 0 \quad OR \quad H_a = \beta_i \neq 0 \quad (4.2)$$

The hypothesis is tested by performing the F-statistic. The F-statistic is defined as (James et al. 2013)

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \quad where;$$
$$TSS = \sum (y_i - \bar{y})^2 \quad and \quad RSS = \sum (y_i - \hat{y}_i)^2 \quad (4.3)$$

In equation 4.3, \hat{y}_i refers to the predicted Y on the i^{th} value of X, \bar{y} refers to the mean of y_i values, y_i refers to the actual i^{th} value of y, n, and p refers to statistical factors describing F-statistics distribution. TSS is the measurement of total variance in response to Y, and RSS is the amount of variation left unexplained after applying regression (James et al. 2013). Here, p value indicates the importance of the feature. If p-value is low after implementing the null hypothesis for a feature, it will indicate a significant impact on output, whereas the feature will have a small or no indication if p-value is high. A process called backward elimination can be applied to implement a null hypothesis to each feature. The process of backward elimination is described below in Figure 4.1

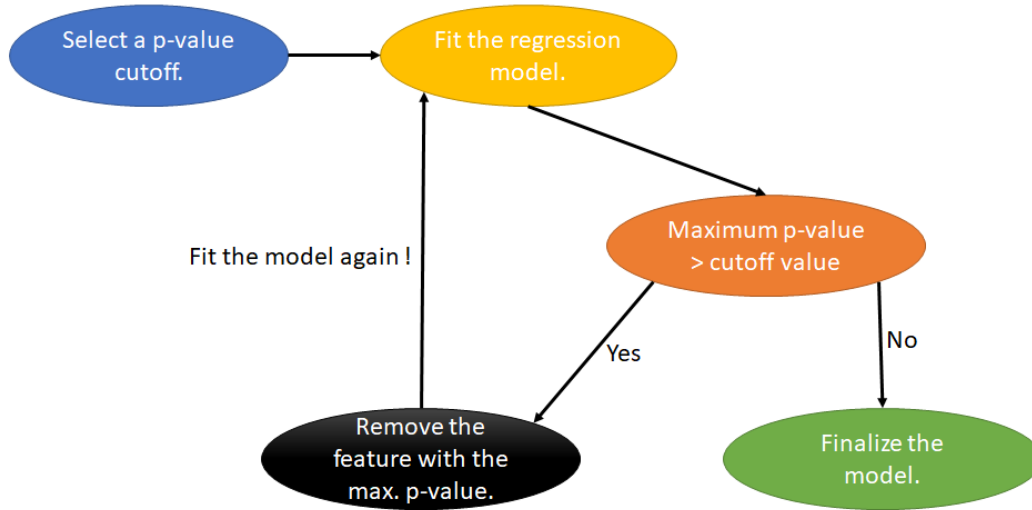


Figure 4.1: The backward elimination flowchart

Even though the multiple linear regression method is robust and can be used for complex regression problems, the multiple linear regression approach is not complex enough to cover the drilling data patterns due to the dynamic and complex nature of drilling.

4.1.1.2 Polynomial Regression

Polynomial regression is defined as adapting a linear regression model to fit a non-linear relation using power features. The polynomial regression can be described as:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * (x_i)^2 + \beta_3 * (x_i)^3 + \epsilon_i \quad (4.4)$$

Polynomial regression can be a robust method to fit a non-linear model. The process to fit a non-linear model is the same as a linear regression model, but it should be noted that polynomial regression models tend to overfit, which is essentially memorizing patterns in training data set and replicating the same results. The overfitting causes the model to be non-generic and results in a model that can only work with the same training data set. This can be avoided by implementing cross-validation methods. The cross-validation methods will be explained later in this chapter. Overfitting can be observed clearly for 300th degree polynomial regression case on Figure 4.2. A linear regression, 2nd degree polynomial regression, and 300th degree polynomial regression models fit the same training data set to observe possible overfitting. For 300th degree polynomial regression case, a model recognizes even insignificant patterns in the data set, which results in the model that

will perform poorly if another data set is used to estimate target values.

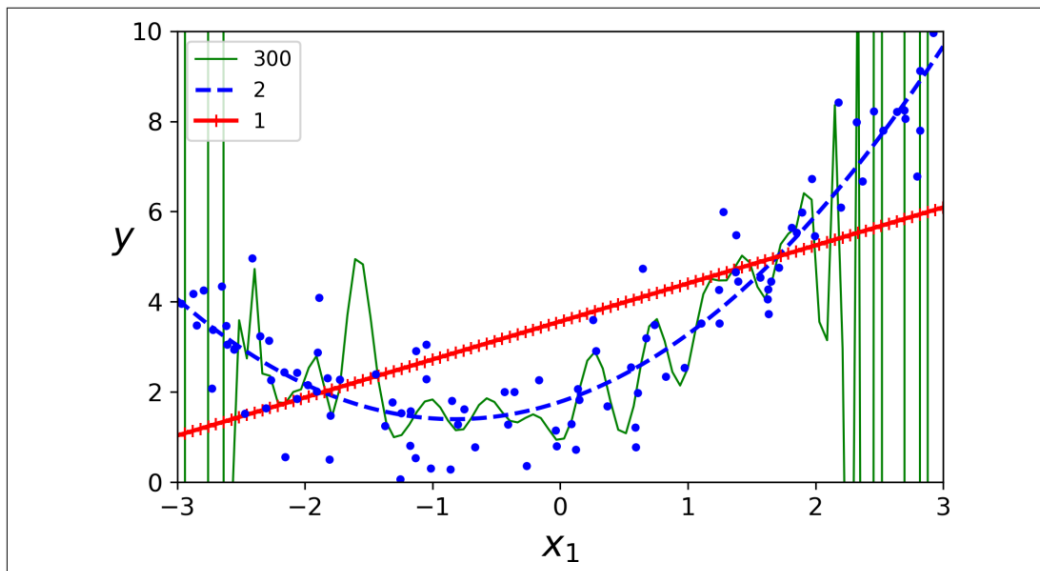


Figure 4.2: Linear Regression, 2nd, 300th degree Polynomial Regression models on the same training dataset (Geron 2019) ©2019 Kiwisoft S.A.S. Published by O’Reilly Media, Inc. Used with permission

4.1.1.3 Support Vector Regression

Support Vector Machines (SVM) is a type of supervised ML algorithm that sorts the data into two categories based on classification differences, and it uses these two categories to identify difference within the dataset. Even though SVM commonly used for classification problems such as image classification, recognition of handwritten characters, it can be a robust and efficient algorithm for regression problems. Support Vector Regression (SVR) is an extension SVM that is used to bring solution to regression problems. Similar to SVM, SVR creates (n-1) number of linear or non-linear hyper-planes based on predefined function or kernel with bounds ϵ distance away from these hyper-planes. SVR estimates target values based on the data points laying outside of margins set at a distance ϵ from hyper-plane, and feeding the model with more data points to the margin placed within the boundaries don’t necessarily impact the estimations (Geron 2019). The mathematical approach to SVR formulation can be best described with geometrical perspective by approximating continuous-valued function for one-dimensional data (Awad and Khanna 2015). The example model of one-dimensional linear SVR is given in Figure 4.3. As it can be observed in

Figure 4.3, possible support vector are determined by data points outside of margin determined by ϵ distance from hyper-plane.

$$y = f(x) = \langle w, x \rangle + b = \sum_{j=1}^M M(w_j x_j + b) \quad \text{where; } y, b \in R \text{ and } x, w \in R^M \quad (4.5)$$

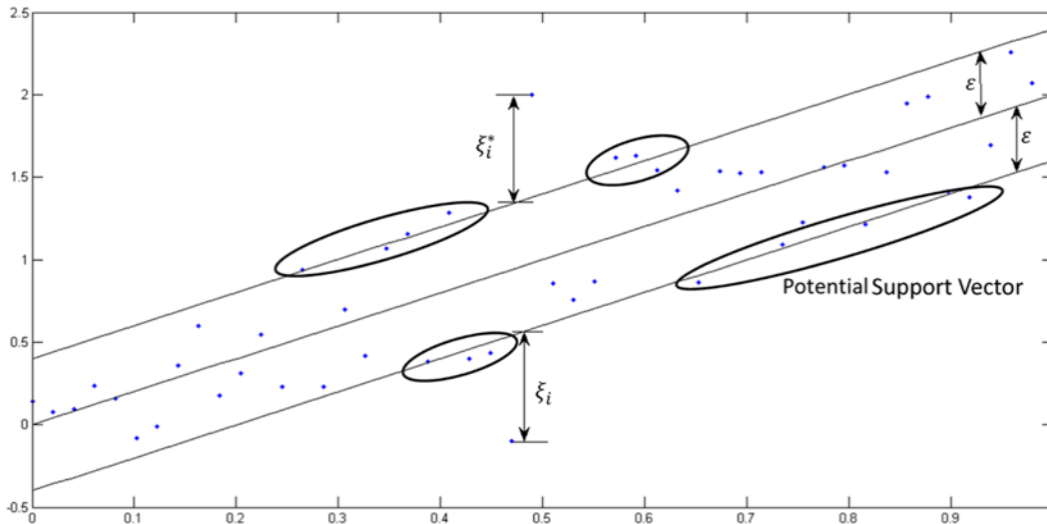


Figure 4.3: One-dimensional linear SVR example (Awad and Khanna 2015)

4.1.1.4 Random Forest Regression

Random forest is a tree-based model that can be a useful method to build efficient regression models. Decision tree and random forest models follow the same splitting rules based on information gained by splitting the training data set. Detailed information about decision tree methods and random forest are provided in Section 2.3 and 2.3.1, respectively. Decision tree models can be a robust solution to complex regression problems. The regression tree models split the training data set into branches (leaves) until the limit data set allows. While training decision tree regression models, data will be split into small batches until reaching the terminal node. The predicted value is estimated based on the mean of the terminal node that the input vector activates (Joshi 2021). Figure 4.4 clearly shows how the data split among leaves. The estimation process of ROP based on WOB and RPM can be observed in the decision tree below. In Figure 4.4, the relation between ROP, WOB, and RPM as data sets is split into small batches by applying various split rules (Joshi

2021). By using the decision tree below, $ROP = 160$ ft/hour can be predicted for $WOB = 5$ klb and $RPM = 150$.

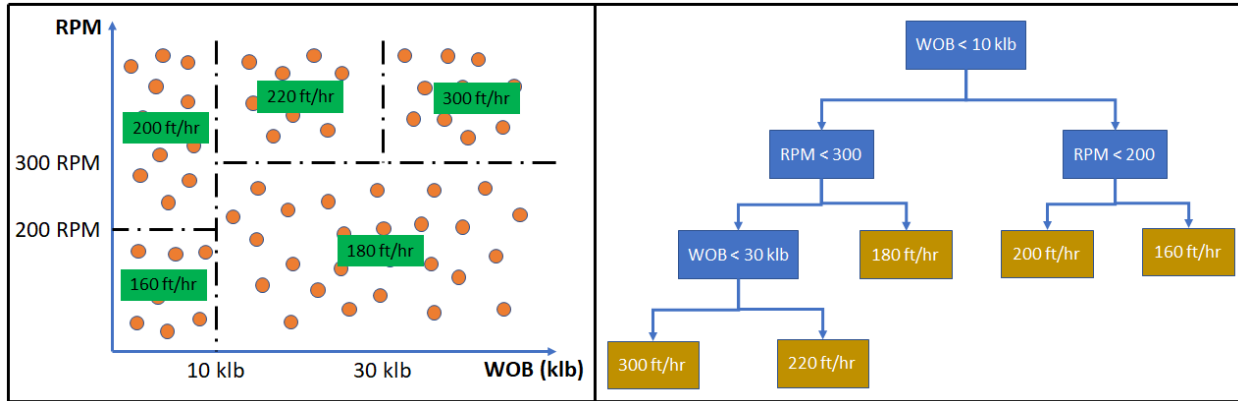


Figure 4.4: (Left) The relation between WOB, ROP, and RPM based on split rules, (Right) The decision tree to predict ROP. Modified from Joshi 2021.

Even though decision tree regression models are powerful tools to build regression models around a vast amount of data, there are several issues regarding their low generalization and high variance. In other words, small changes in the data set can drastically impact predictions. A group of trees can be built to reduce this impact, and collective decisions made on each tree can be combined to provide a final prediction. This method is called bagging or bootstrapping, and it is explained in detail in Section 2.3. Bagging or bootstrapping is using the same batch of data set to train different predictors to build a group of weak predictors instead of a single strong predictor. The bagging can significantly reduce the variance in predictions as each prediction delivered from a group of predictors is taken into account and averaged to estimate the final output. Random forest implements bagging as part of its function and splits each predictor's data set based on randomly sampled input features.

4.2 Possible Problems and Solutions

This section presents possible problems regarding the implementation of machine learning models for regression problems. Generally, complex regression problems can be solved by using machine learning models, but these models require hyper-parameter tuning to prevent common problems such as low accuracy, underfitting, overfitting due to bias, and variance errors. The difference

between predicted and actual output value is called error, and any supervised machine learning algorithm is compromised to noise, bias, and variance error. Even though the noise error, aka irreducible error, can not be eliminated, bias and variance error can be minimized to a certain degree.

4.2.1 Bias and Variance

Bias is defined as a generalization error caused by incorrect assumptions, such as modeling 4th polynomial regression as linear regression (Geron 2019). Also, bias can be defined as the difference between predicted and actual values. The model does not learn enough to make accurate predictions if it has a high bias error, and eventually, it will cause an underfitting of the training data.

Contrary to bias, Variance error is caused by learning from training data very well. Variance error is defined as excess sensitivity of the model to small patterns in the training data set (Geron 2019). The high variance error is generally caused by giving a high degree of freedom to a model since a model with the freedom to choose a degree of fit has a tendency to fit a higher polynomial regression. The high variance error eventually causes overfitting of the training data.

4.2.2 Underfitting and Overfitting

As mentioned in the previous section, underfitting and overfitting are the results of high bias and high variance.

Overfitting is a phenomenon of well-performing model on the same training data set, but the model can't generalize or performs poorly on different data sets (Geron 2019). A real-life example of overfitting can be generalizing an attitude of a bad taxi driver and assuming that every taxi driver will act in the same manner (Geron 2019). Overfitting is caused by high variance error, and it can be described as the significant impact of insignificant variations or patterns in the training data set on the overall learning of the model. An example of overfitting can be observed in Figure 4.5. The model assumed that the training data is more complex due to small variations in the data set, which led to building a high-degree polynomial regression model instead of a simple linear regression model. Some ways to avoid overfitting are re-processing the training data to fix data errors and remove outliers, using a bigger training data set, simplifying the model by dropping the features that have small or no impact on the learning process (Geron 2019).

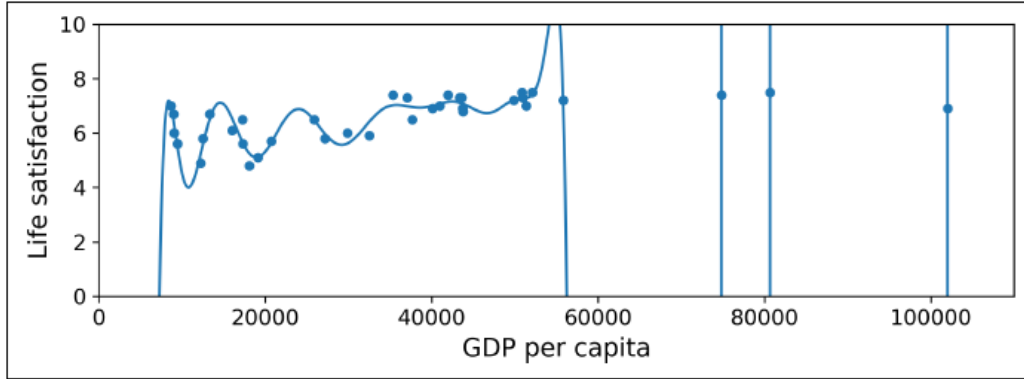


Figure 4.5: An Example of Overfitting the training data (Geron 2019)©2019 Kiwisoft S.A.S. Published by O’Reilly Media, Inc. Used with permission

As mentioned, underfitting is the opposite of overfitting, and it is caused by high bias error. Underfitting can be defined as the model’s lack of complexity due to incorrect assumptions (Geron 2019). Commonly, linear regression models tend to underfit the training data. An example of underfitting can be observed in Figure 4.2. The linear regression shown in the red line in Figure 4.2 indicates that the simple regression model does not cover the variations in the training data enough to make accurate predictions. Some of the ways to avoid underfitting are choosing a more complex and robust model with more features, reducing the model’s limitations by tuning hyper-parameters, and using better features to train the model (Geron 2019).

4.2.3 Bias and Variance Trade-off

As explained in the previous sections above, bias and variance introduce opposite types of error to a regression model. The high variance causes overfitting, whereas the high bias causes underfitting. A trade-off is required to develop the best approach to limit both bias and variance as an increase in variance generates a more complex model, leading to high bias. The relation between bias and variance is visualized and provided in Figure 4.6. There are several methods to achieve optimum model complexity. These methods are cross-validation, bagging, removing outlier data, and hyper-parameter tuning. While building Random Forest regression model for this study, bagging, cross-validation, and hyper-parameter tuning has been utilized to optimize the model. As mentioned in Section 4.1.1.4, bagging is a built-in function of the Random Forest regression model, and it reduces variance error, which is a common problem of decision-tree regression models. Also,

the Random Forest regression model offers flexibility to tune hyper-parameters, which helps reduce bias in the model.

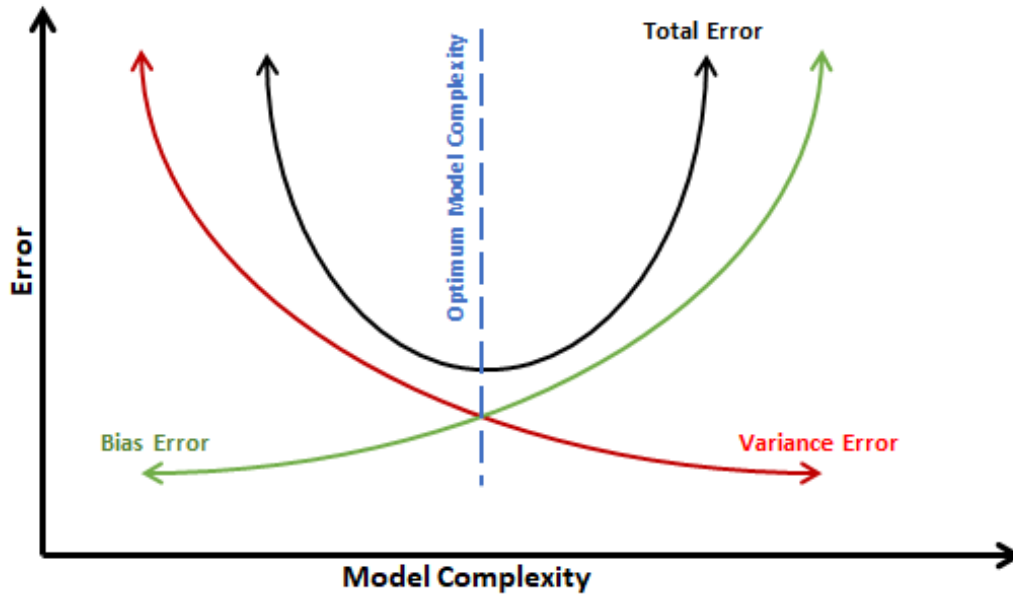


Figure 4.6: Bias - Variance Trade-off

4.3 Hyper Parameter Tuning

Random Forest regression model builds regression trees based on predetermined parameters. There are 17 different parameters that can be optimized to achieve optimum accuracy while avoiding overfitting the training data set. The optimization of these parameters is called hyper-parameter tuning. These can be achieved by using different built-in functions packed in the scikit-learn library in Python (Pedregosa et al. 2011). Some of these methods from scikit-learn library are Cross-Validation, K-fold cross-validation Randomized-Search, and Grid-Search (Pedregosa et al. 2011). The most common method to optimize the parameters is K-fold cross-validation, and it can be integrated with both Randomized-Search and Grid-Search by using the built-in function from scikit-learn library in Python (Pedregosa et al. 2011).

K-fold Cross-validation is a verification method to evaluate the model's generalization capability. It splits the training data set into small mutually exclusive data batches and uses them to test the model while training for overfitting. K-fold Cross-validation can also be utilized to check the

accuracy of predictions. While applying K-fold Cross-validation, the training data set is sub-sampled to have k mutually exclusive data batches, and the model is tested with one of the k data batches on each iteration, whereas remaining (k-1) batches are used to fit a model. K-fold Cross-validation implements an iterative approach to test the accuracy of predictions, so every data batch is used as both train and test subset. The implementation of K-fold cross-validation is an example presented in Figure 4.7 to clarify the concept of cross-validation.

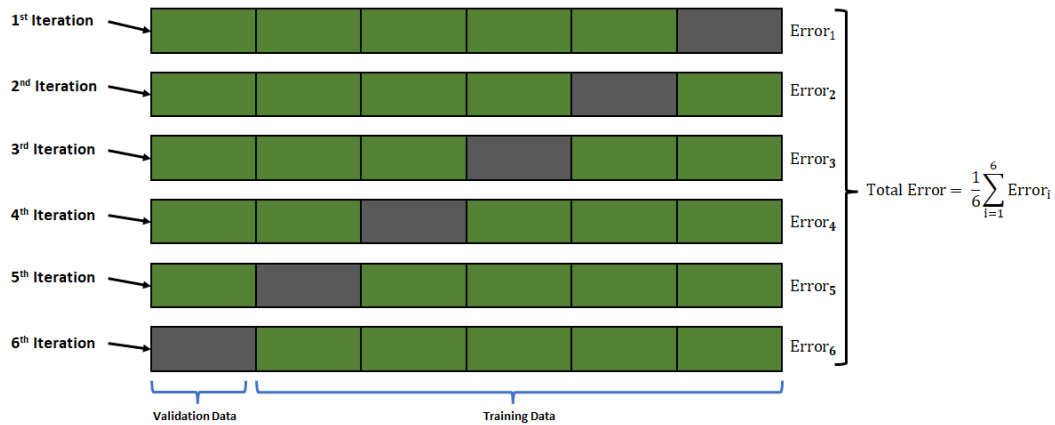


Figure 4.7: K-fold Cross Validation Example

Both Randomized-Search and Grid-Search are optimization tools developed as a built-in function in scikit-learn library (Pedregosa et al. 2011). Both methods use an iterative approach to find best-fitting parameters as they continuously fit the model based on a preset list of parameters. The advantage of using these methods is implementing an iterative approach to try out combinations of the listed parameters to achieve optimum accuracy in predictions. The fundamental difference between Randomized-Search and Grid-Search is how the combinations of parameters are selected. While using Randomized-Search, the parameters are selected randomly from a range of numbers to fit that individual instance, whereas these parameters are selected from a list of numbers preset by the author while using Grid-Search. These methods are powerful tools to improve the model's generalization as K-fold Cross-Validation is used for each fit. The implementation of these methods is similar and is presented in Figure 4.8 as a flowchart.

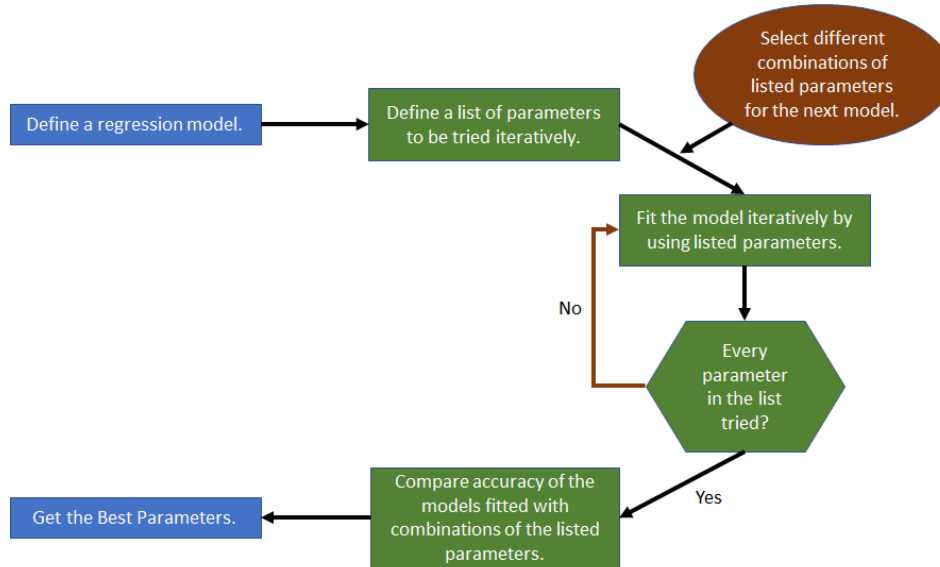


Figure 4.8: RandomizedSearchCV-GridSearchCV Steps of Implementation

4.4 Evaluating Performance of the Models

The performance parameters of a regression model can be calculated by using several methods. The process of validating a model is simply estimating the performance of the trained model based on the error margin between actual and predicted values. There are several statistical approaches to measure the accuracy of a model. For regression models, the most common methods to evaluate the accuracy are Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) (Geron 2019). This section evaluated three main methods, including RMSE and MAE (Swalin 2018).

Root Mean Squared Error (RMSE): Root mean squared error is the most common method to estimate the accuracy of a regression model. For a regression model, the equation below is used to calculate RMSE. In Equation 4.6, y_i refers to the actual value at i^{th} observation, \hat{y}_i refers to the predicted value at i^{th} observation, where n refers to the number of predicted values for each observation. The values of RMSE will be small, if the predicted values are close to the actual values.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.6)$$

Mean Absolute Error (MAE): Mean absolute error is fairly similar to RMSE, and it can be calculated using Equation 4.7. The fundamental difference between RMSE and MAE is that MAE

treats errors equally, whereas RMSE penalizes significant errors by square.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.7)$$

R^2 and Adjusted R^2 : R^2 can be defined as estimating variations within the predictions based on features fed into the model. The range of R^2 is between 0 and 1. The higher R^2 value represents accurate predictions and less variance within the prediction compared to actual variance within the training dataset. The R^2 can be calculated by using Equation 4.8. On the other hand, adjusted R^2 considers the number of features fed to the model by considering how useful the feature is. If a feature with a significant impact is fed to the model, adjusted R^2 will increase, whereas it will decrease if a feature with no impact is fed to the model (Swalin 2018).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.8)$$

The last step of evaluating these models was choosing the most efficient method. R^2 and Adjusted R^2 were powerful methods to estimate the accuracy, but Ford (2015) argues against using this method based on the initial arguments proposed by Shalizi (2015). Ford (2015) argues that R^2 can estimate the accuracy of the model's accuracy as close to 1 even though the fitted model is completely wrong if the variance in the training data set is high.

Considering a high variance and noise in the drilling data, R^2 is considered the least preferred method to measure the model's accuracy. RMSE and MAE were suitable options to measure the accuracy of UCS estimations. Hence, both RMSE and MAE are used as the performance evaluation method in this study.

4.5 Final Architecture of the Model

As mentioned before, the final architecture of the regression algorithm trained to estimate UCS from drilling parameters is explained in this section. The process of training the model is described and provided in Figure 4.9. The overall objective of the algorithm is to train a robust Random Forest regression model while optimizing the accuracy of estimation without overfitting the drilling data.

The training process is completed in three main steps, and at each step, a different function is used. These steps prepare the training data set, optimize the model, and evaluate the final model's performance.

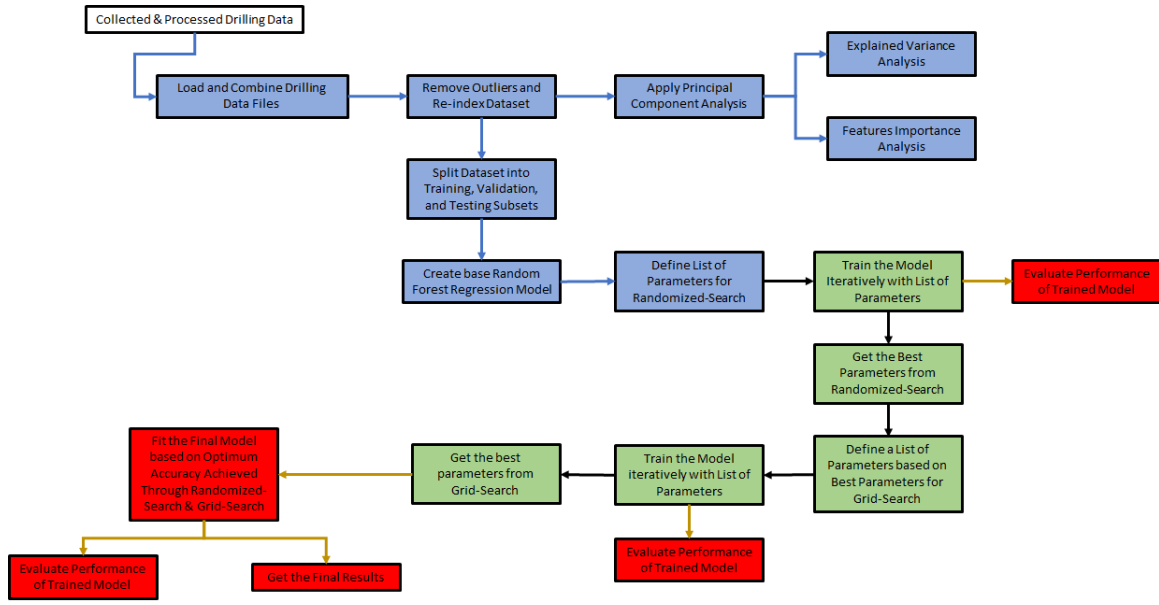


Figure 4.9: The Final Architecture of prediction model.

- Preparation of the training dataset:** The overall goal of this step is to implement Principal Component Analysis to prepare the training data set, gain insights into the correlation between features, and estimate explained variance in the data set by each principal component. Also, the data set is split into three subsets. The first set was testing, aka blind test data set. The blind test data set was 10% of the original data set, and it was never shown to the model until the final model was fit. The main reason to split the testing data set was to ensure that the model never learned from this sub-set of original data. Hence, the testing data set can be used to check if the final model is overfitted. The remaining data set is split into two subsets, training and validation data. The training and validation subsets consisted of 80% and 20% of the remaining data set.
- Optimization of the Model:** In this step, the best parameters are searched by using Randomized-Search and Grid-Search. These methods fit the model numerous times continuously by using a preset list of parameters. Each fit uses the K-fold cross-validation method to avoid overfitting. At the end of each iterative method, a combination of the best parameters is chosen. The model is fit one more time based on these best parameters, and the accuracy of predictions is evaluated before initiating the next step.

- **Training and Evaluating accuracy of the model:** This is the final step of the model. After implementing Randomized-Search and Grid-Search, the best parameters for the final model are selected. Based on these parameters, the final model is fit, and the accuracy of predictions is evaluated. Also, the testing subset split from the original data set is used to evaluate the performance of the final model. This testing subset is called a blind testing data set, and it is an essential step that helps identify overfitted training data.

4.6 Discussion on Challenges and Changes

The path of this study changed with time due to the particular problem of high accuracy in predictions of Random Forest regression model trained by the principal components of drilling parameters. The accuracy of UCS predictions was as high as 97% on the blind testing data set, which indicated that the model is overfitted. These results were highly unexpected as bagging and bootstrapping were used while training each decision tree of Random Forest regression algorithm to reduce variance error. In addition to this, K-fold cross-validation was used on each model fit while tuning hyper-parameters to reduce both bias and variance error. Even though the initial exploratory data analysis on drilling data indicated a lack of variability in UCS as only seven different UCS values were present, however, this issue was never assumed to cause an overfitting problem while building a regression model. After realizing the overfitting problem, the model was fitted numerous times with a different list of parameters fed into hyper-parameter tuning tools Random-Search and Grid-Search, but the model was performing poorly, and the generalization of the model was low. To achieve the study's overall objective, the following hypotheses are tested by conducting comprehensive research to indicate the root cause of the overfitting problem.

- A possible linear correlation between MSE and UCS can lead to perfect predictions.
- A loss of valuable information due to the implementation of PCA can cause overfitting.
- A lack of variability on UCS can reduce the variability of predictions, which can cause overfitting.

Pearson correlation coefficient between MSE and UCS is calculated to test the first hypothesis. Pearson correlation coefficient is defined as a measurement of linear correlation between two

variables. The coefficient value ranges from -1 to 1, and perfect positive correlation is indicated by 1, while perfect negative correlation is indicated by -1. Pearson coefficient between MSE and UCS is calculated as -0.189, which indicates no correlation. The first hypothesis is tested and eliminated. The second hypothesis was tested by going through the implementation of PCA on each data subset. These data subsets are data files in feather format. Initially, the data set collected by Joshi (2021) was provided in four different feather files. These data files were initially divided as it was computationally expensive to load the complete data set. The initial feature importance and explained variance analyses are conducted on these four subsets separately. The explained variance retained by the first four principal components was over the threshold value of 85% for all subsets, as observed in Figure 4.10. This value was set as a target explained variance retained for this study to use principal components of drilling parameters as a training data set.

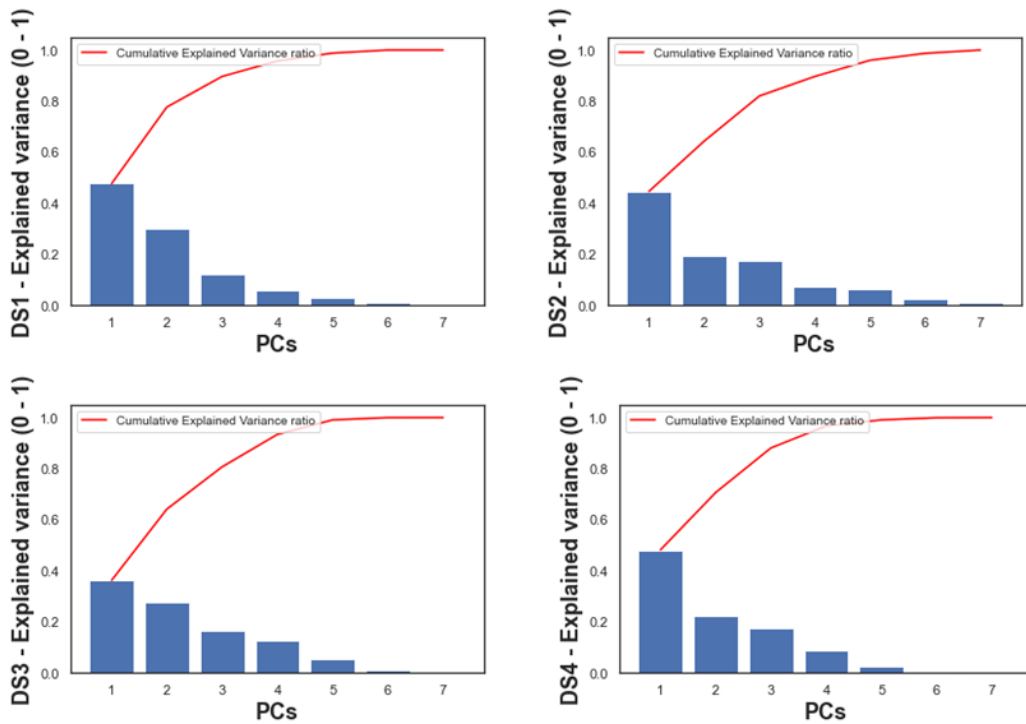


Figure 4.10: The Explained Variance Retained by PCs for Datasets 1 to 4.

After analyzing retained explained variance percentages, feature importance analysis is conducted to indicate the influence of variables on each principal component. The feature matrix for each subset is visualized to conduct a feature importance analysis. As a result of observations from

Figure 4.11 and Figure 4.12, the feature importance analysis indicated that there was small or no influence of variables on principal components 5, 6, and 7. The initial observations from explained variance and feature importance analyses indicated that using principal components of drilling parameters did not lead to information loss. With these results, theoretically, a similar result should have been observed on the complete data set. However, the observations in the results of feature importance and explained variance were significantly different. The observations from Figure 3.5 and Figure 3.7 indicated that explained variance ratio by four principal components was lower than the threshold value of 85% set for this study. Also, feature importance analysis indicated a significant influence of variables on Principal Component 7. These results indicated an information loss introduced by implementing PCA, leading to an overfitting problem. As a result, PCA has been taken out of the algorithm to observe changes in the accuracy of predictions. Unfortunately, the removal of PCA from the algorithm completely did not solve the problem of high accuracy, as the accuracy of fitted models was approximately 98%. With this result, the second hypothesis is tested and eliminated. However, PCA is not added back to the algorithm as information loss on removing Principal Component 7 is observed in the analyses.

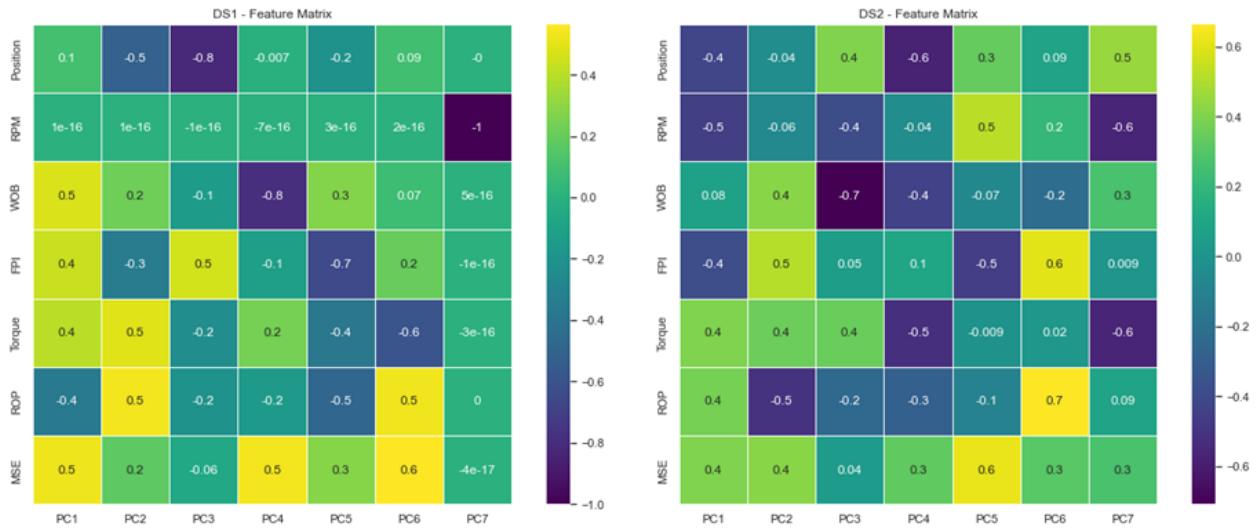


Figure 4.11: The Feature Matrix of Datasets 1 - 2.

Testing the last hypothesis was not as straightforward as other hypotheses since variability should be introduced to target outputs, and collecting a new data set were impossible. In addition, the raw data set was not accessible since the algorithm built by Joshi (2021) was working as a whole

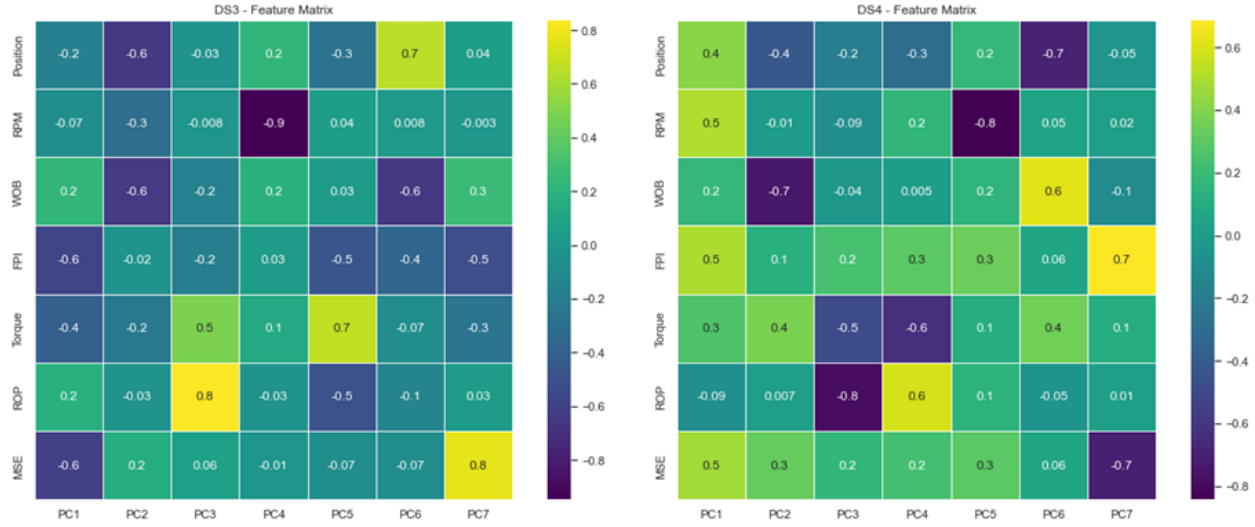


Figure 4.12: The Feature Matrix of Datasets 3 - 4.

system that takes collected drilling data as an input and predicts Torque values after classifying the raw data as drilling and non-drilling data. Also, high accuracy in UCS predictions is observed by Joshi (2021), as it is stated that the UCS regression model predicted UCS values for analog and cryogenic samples with less than a 2.5% error margin. However, the issue of overfitting UCS did not introduce a significant generalization problem as the purpose of the final algorithm built by Joshi (2021) was to indicate drilling dysfunctions while drilling, which is essentially a classification problem. The finding from a previous study conducted by this data set shows that another approach should be developed to indicate if the regression algorithm built for this study is working or not. Potential performance changes in the Multi-Output Random Forest algorithm are studied with the idea of increasing variability by introducing another target value to the model. The study conducted by Linusson (2013) on Multi-output Random Forests indicates that the performance of multi-output random forest models should be similar to or the same as single output models. Then, one of the variables should be moved to target to increase variability within the output values. MSE is selected as a variable to move to the target values. As a result of moving MSE to target values, approximately a 5-10% decrease in average accuracy was achieved on every fitted model. With this, the third hypothesis is tested, and the performance improvement indicated that the lack of variability in UCS data was leading to perfect predictions and causing the model to overfit the training data.

4.7 Potential Implementation of the Model for Field Applications

Throughout this study, the improvements that can be introduced by implementing data-driven solutions to the field applications are mentioned and discussed based on previous studies conducted with similar intent. This section discusses the potential implementation of the regression model built step by step. As mentioned in previous sections, the regression model built aims to estimate UCS from the drilling parameters instantaneously and avoid potential wellbore stability problems and drilling accidents by providing changes in UCS while drilling. The model should be re-trained with necessary data collected from the field where it will be used as a UCS prediction tool since data-driven solutions are as unique as the data fed into the model. The potential development of the UCS prediction tool by using this model for field applications is studied, and necessary steps are provided in Figure 4.13. The initial implementation step will be using available drilling and

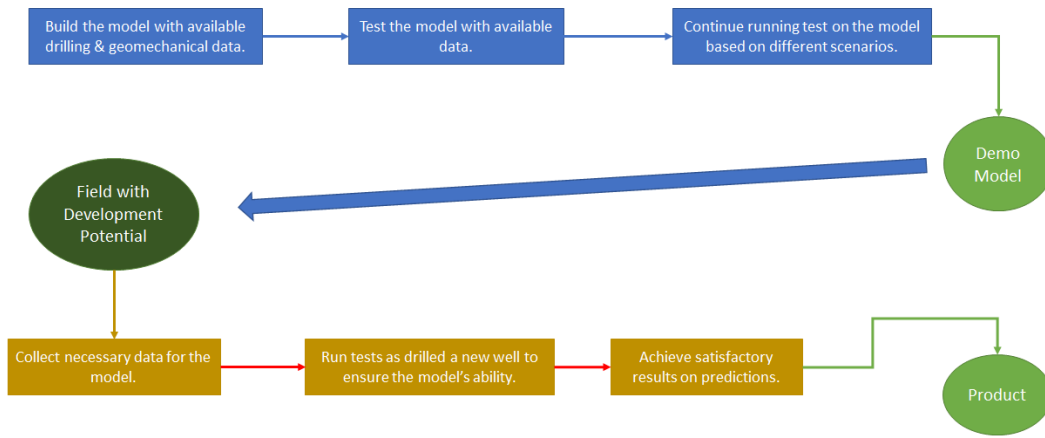


Figure 4.13: Potential Implementation on Field Application Workflow

geomechanical data from previous wells drilled to train the model and test the model's performance. Then, the model can be tested based on different scenarios to indicate the model's reaction to real-life drilling operations (i.e., drilling break, high dogleg severity). After running these tests, a demo product can be produced using the model if satisfactory test results are obtained. The demo product can be tested in a field with development potential by training the model after collecting necessary drilling and geomechanical from 1st exploration well. The model's ability to predict UCS while drilling can be tested for every appraisal well drilled and the additional data collected can be fed to the model to improve its ability to indicate previous problems faced while drilling. A final

product can be achieved if the test results indicate satisfactory results on the model's veracity and versatility. The final product can be used as a robust tool to predict UCS from drilling parameters instantaneously while drilling development wells on the field.

CHAPTER 5

RESULTS AND TECHNICAL EVALUATION

The analysis was conducted on data set, and the findings are previously mentioned in Chapter 4. The final architecture of the algorithm is given, and possible problems that might occur while training RF regression algorithm are explained in detail in Chapter 4. The final and third step of the algorithm is described in this chapter. In addition, the overall results of models fitted by using different approaches are provided. The changes in the algorithm as a consequence of findings comprehensively discussed in the previous chapter are briefly summarized. A problem of overfitting indicated due to a lack of variability within UCS data for each model is evaluated by visualizing predicted and actual UCS values. Finally, the final model fit by using UCS and MSE as target values, and the results are visualized and evaluated in this section.

5.1 The development phase of the algorithm

As mentioned before, the RF algorithm built to predict UCS is fit numerous times, and hyper-parameter tuning tools are implemented at each fit. Initially, the model was fitted numerous times with only 1% of the complete data set to gain more knowledge about RF regression model and achieve the overall objective of building a robust algorithm. These subsets are sampled randomly based on the distribution within the data set. A five K-fold cross validation method is applied to reduce variance error at each fit. These initial tests on the algorithm helped to eliminate minor errors in the code.

5.1.1 Initial Results

While conducting this study, the original plan was to use principal components of drilling parameters to train the RF regression model. As mentioned earlier, 85% explained variance retained principal components set as a threshold value. Initially, only 5% of data sampled from each file are fed into the PCA algorithm to transform into principal components. The results acquired by implementing PCA before feeding the training data into the algorithm were promising, and using only fraction of the complete data set helped avoid minor errors. In addition to hyper-parameter

tuning, the percentage of validation data split and implementation of K-fold cross-validation are tested in these initial trials. The performance of each fit is evaluated with MAE, MSE, RMSE, and MAPE. The overall accuracy is calculated from MAPE. The results are given in Table 5.1. After running through the performance evaluation algorithm, it is clearly indicated that K-fold cross-validation reduced variance error by preventing overfitting.

Table 5.1: Initial models fitted with sub datasets after implementing PCA

Number of PCs	K-Fold CV	Validation Dataset Size	MAE	MSE	RMSE	MAPE	Overall Accuracy
3	No CV	20%	0.17	1.15	1.07	0.23	99.77%
2	No CV	20%	3.61	103.31	10.16	4.29	95.71%
2	No CV	30%	3.61	103.46	10.17	4.28	95.72%
3	No CV	30%	0.17	1.2	1.09	0.25	99.75%
5	No CV	30%	0.046	0.018	0.13	0.03	99.97%
3	No CV	30%	0.18	1.24	1.11	0.25	99.75%
5	3 K-Fold	20%	20.33	1042	32.38	28.34	71.66%
5	4 K-Fold	30%	16.43	682.9	26.13	21.9	78.10%

Later, PCA was implemented on each data file provided, and exploratory analysis on these data sets was promising since explained variance ratio retained by four principal components were higher than 85%, and the influence of variables on principal components was low, as discussed in detail Chapter 4. After implementing PCA, seven drilling parameters are replaced with four principal components, as four PCs explained most of the data variance present in the complete dataset. After these steps, PCs are fed to the algorithm as a training dataset. The hyper-parameters are searched using Randomized - Search and Grid-Search while fitting the model numerous times to indicate the best possible fit. The final parameters used while fitting the last mode is given in Table 5.2. The followings refer to the terms indicated in Table 5.2.

- Number of estimators: Number of trees created by using training data, and the final prediction is made by the contribution of each of these estimators.
- Maximum features: Decision criteria for maximum amount of features considered while splitting node.

- Maximum depth: Maximum number of levels created while building each decision tree (estimator).
- Minimum samples split: Minimum number of data points placed in a node before it splits.
- Minimum sample leaf: Minimum number of data points to be kept in each leaf node.
- Bootstrap: Method to sample data points.

After fitting the model, MAPE is used to indicate model performance, and 98% accuracy in UCS predictions was observed. Extremely accurate predictions were a clear indication of overfitting.

Table 5.2: Hyper-parameters of the model fitted with four PCs

Hyper - Parameter	Selected Parameter
Number of estimators	35
Maximum features	Square root
Maximum depth	40
Minimum sample split	9
Minimum sample leaf	2
Bootstrap	True

The UCS predictions are visualized to observe possible problems introduced by the lack of variability in UCS data. Initially, the results from a model trained with one of the subsets are visualized to observe how the lack of variability within UCS data affects the final visuals. As expected, raw predictions collected from the model could not be visualized properly since the predictions were too noisy. The moving average method is used to smooth the results. The moving average can be defined as a statistical method for smoothing noisy predictions by continuously taking the average values of the range defined. The advantage of moving average is keeping the impact of each variable taken into consideration while continuously calculating the mean of a given range of numbers. For example, the moving average of integers from 1 through 5 in the range of 2 would be 1.5, 2.5, 3.5, and 4.5. A closer match between actual and predicted values was observed after using the moving average. The actual and predicted UCS values from the model trained with a subset is plotted after applying a moving average with a range of 20, 50, and 100, given in Figure 5.1.

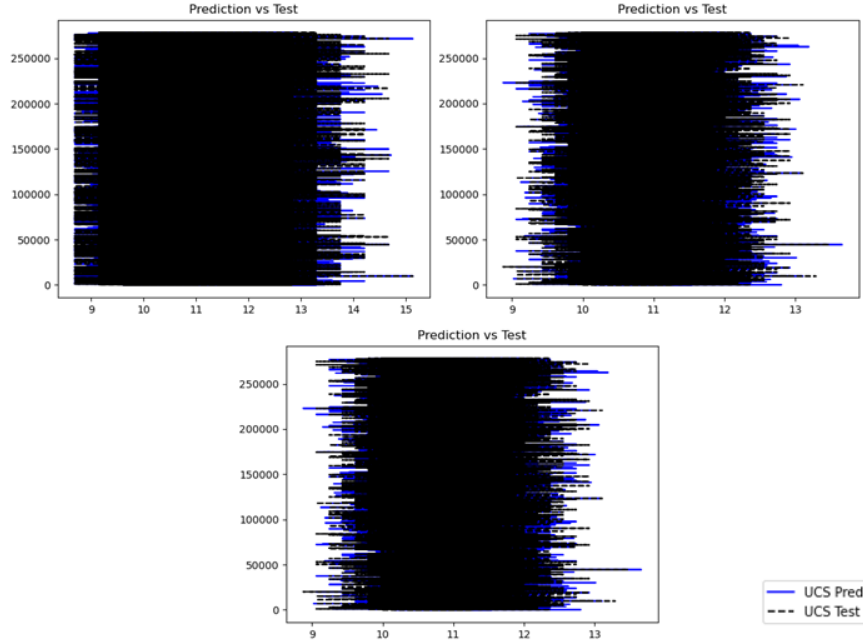


Figure 5.1: Comparison between predicted and actual UCS with 20, 50, 100 moving averages.

The significant noise in predictions is observed from the model trained with 4 PCs of complete data set also indicated a presence within the predictions, and a higher range of moving average should be used to observe a proper correlation between prediction and testing data. Then, a 10,000 moving average is applied to 5+ million predictions, and the observations on results were more promising to indicate almost a perfect match between actual and predicted UCS values, as observed in Figure 5.2. These results show a high variance error and indicate overfitting. As mentioned in Chapter 4, PCA has been removed from the final algorithm.

5.1.2 Models Fitted without PCA

To test the 2nd hypothesis discussed in detail in 4, PCA is removed from the main algorithm. Later, PCA is used to conduct explained variance and feature importance analysis. After removing PCA, the number of variables fed to the algorithm is changed to seven; the overall objective was to predict a single target value, UCS. The variables fed to the algorithm are Depth, RPM, WOB, FPI, Torque, ROP, and MSE. Different cases are considered by varying training data set size and number of iterations.

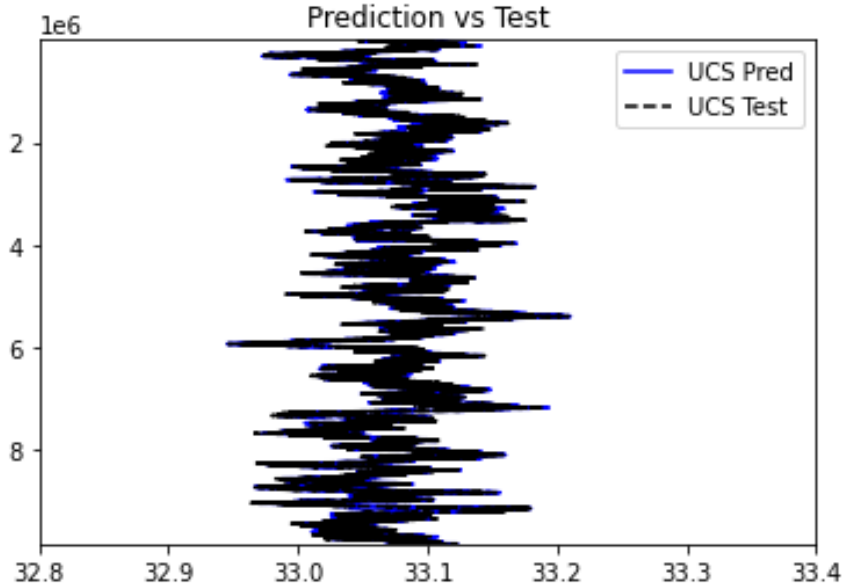


Figure 5.2: Predicted and actual UCS data with 10000 moving average - Predicted Values (Blue), Actual Values(Black).

Initially, RMSE was planned to evaluate the model's performance, but MAPE has replaced it since RMSE values were incredibly low, and it was impossible to indicate any difference between models. MAPE values indicated a performance change between fitted models in the third decimal point. The motivation for removing PCA was to study a possible information loss by using only four principal components.

The second hypothesis was tested by removing PCA to indicate if initial information loss caused the overfitting observed in previous results. Similar to previous graphs describing predicted and tested UCS versus depth created for the model fitted without PCA. Two different cases are evaluated. For the 1st case, only 50% of the complete data set is fitted through 280 iterations with four K-fold cross-validation. For the 2nd case, the entire data set is fitted with five K-fold cross-validation and 250 iterations. Unfortunately, the overall results did not change as UCS accuracy in predictions was approximately 98%. This outcome proved that the information loss through the implementation of PCA was not causing the overfitting problem. For 2nd case, the comparison between UCS values from testing data and predicted UCS values is completed by observing the results in Figure 5.3.

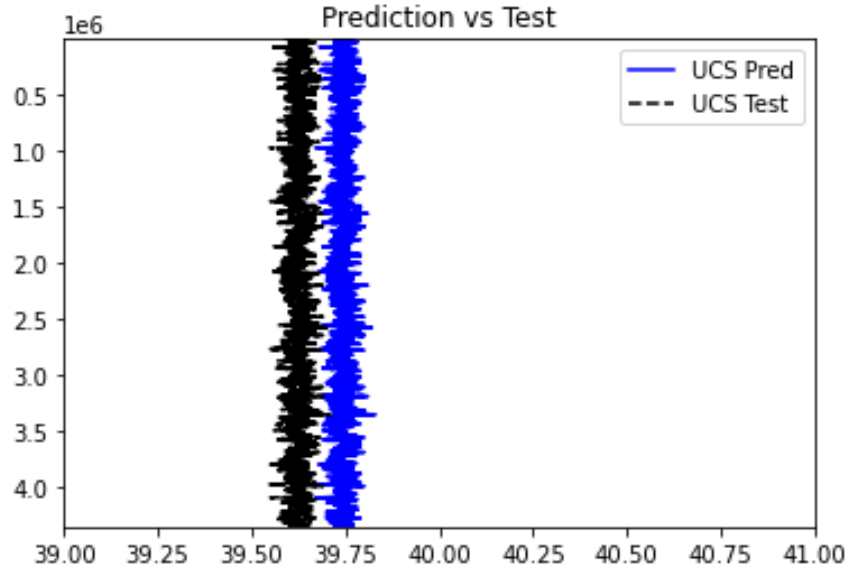


Figure 5.3: Comparison between prediction and test UCS data with 10000 moving average.

5.1.3 Multi-output Random Forest Model

To test the 3rd hypothesis, one of the variables moved to target values. In this hypothesis, the theory of how introducing a variation to target values would change the prediction accuracy. From variables, MSE is moved to target values. Then the model is fitted with six variables (Depth, RPM, WOB, PFI, Torque, ROP) and two target values (UCS, MSE). To the hypothesis, the model fit four times based on different cases. A similar performance was expected as a multi-output random forest should perform similar to a single output (Linusson 2013). In these cases, the number of iterations and percentage of data used to fit the model is varied. These cases are given below in Table 5.3. Based on these findings, it was decided to fit the model with only 100 iterations as the number of iterations was not affecting prediction accuracy significantly. Hyper-parameter tuning tools are used to search for the best model by using K-Fold cross-validation 100 times. The final parameters decided to use while fitting the final model are presented in Table 5.4.

After fitting the model for each case, model prediction accuracy is calculated using MAPE. The prediction of accuracy was not sensitive to the number of iterations, but it increased with the increasing number of data points used to train the model. This finding indicated a lack of variability within the data set even though two target values were used to increase variability. The prediction accuracy of each case evaluated with train, validation, and test sets are given in Table 5.5.

Table 5.3: The parameters used for Final Cases

Case	# of iterations	K-Fold CV	Amount of Data (%)
1	30	5	10%
2	100	5	20%
3	200	5	20%
4	100	5	100%

Table 5.4: Hyper-parameters of the model fitted with six features and two target outputs

Hyper - Parameter	Selected Parameter
Number of estimators	50
Maximum features	Square root
Maximum depth	40
Minimum sample split	5
Minimum sample leaf	3
Bootstrap	True

Table 5.5: The prediction accuracy for each model fit

Case	Train Accuracy	Validation Accuracy	Test Accuracy
1	0.9262	0.9087	0.9067
2	0.9342	0.9112	0.9091
3	0.9379	0.9161	0.9152
4	0.9612	0.9582	0.9431

The predicted UCS values are compared with actual values in Figure 5.4. This comparison was completed after applying a 10,000 moving average to actual and predicted UCS as a combination of lack of variability within UCS and 5+ million data points made it impossible to visualize these values on a scatter plot. The perfect match between predicted and actual UCS was expected as the average prediction accuracy of MSE and UCS is 94%. Since moving average had to be applied to see a clear match between actual predicted UCS, prediction accuracy is taken as an indication of model performance. 10,000 moving average on UCS and MSE predictions from the final model is applied. The comparison of predicted and actual MSE is plotted and presented in Figure 5.5. Also, the actual versus predicted values are visualized using a scatter plot after applying a 10,000 moving average and presented in Figure 5.6.

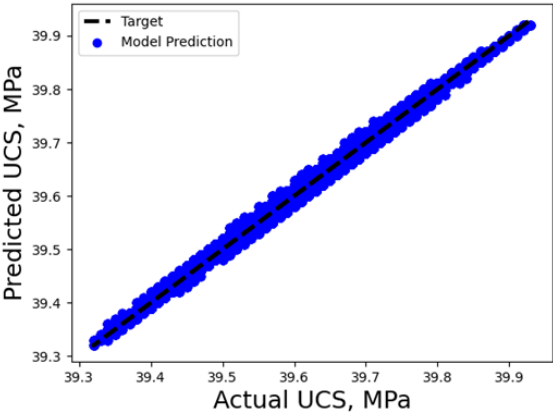


Figure 5.4: Target UCS (Black) vs Predicted UCS(Blue).

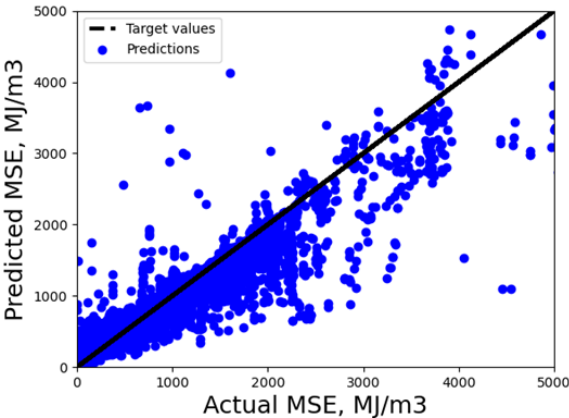


Figure 5.5: Actual and Predicted MSE with the Final Model.

Prediction vs Test (Depth)

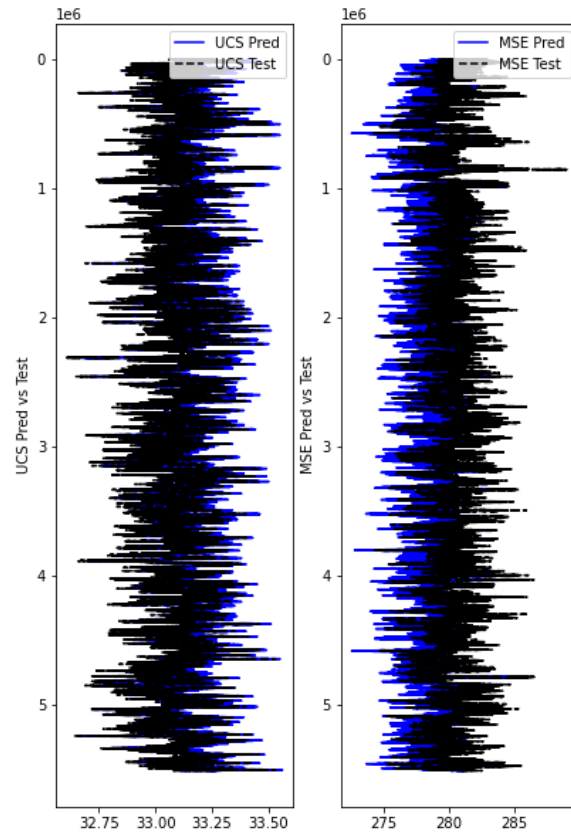


Figure 5.6: Predicted vs Actual UCS and MSE values after applying 10,000 moving average.

CHAPTER 6

SUMMARY AND CONCLUSIONS

The objective of this work was to develop a complex machine learning algorithm by using Random Forest Regression method to process high-frequency drilling data and train a model that can estimate UCS instantaneously from drilling parameters. The importance of gaining knowledge about subsurface conditions and geomechanical parameters is one of the key criteria for efficient drilling operations. Current methods of estimating geomechanical parameters are time-consuming and expensive. These constraints motivate the oil and gas industry to seek much more efficient methods to indicate subsurface conditions and geomechanical parameters. A decrease in the cost of computational power and robust implementation of data-driven solutions in other industries encourages the oil and gas industry to search for more efficient solutions to solve common problems through data collection. By predicting UCS while drilling, the model will allow to show changes in UCS in different zones and give a key input to indicate potential wellbore stability problems.

Initially, this study was proposed to be conducted with a different data set and core samples collected by using a coring rig located at the Edgar mine. The coring rig is equipped with after-market sensors to collect RPM, ROP, WOB, and Torque data. The UCS and Young's Modulus of formation drilled through with the coring rig was planned to be measured by conducting a laboratory experiment at the Colorado School of Mines campus using the MTS Rock Mechanics System. Even though the unique drilling data was collected, UCS data could not be collected as the project milestones did not match this study. A similar data set, collected Joshi (2021), is utilized for this study as it was purposed to build a similar stack of machine learning and deep learning algorithms. The data set collected by Joshi (2021) was re-purposed to achieve the main objective of building a regression algorithm to predict UCS from drilling parameters. The main objectives of this thesis are:

- Develop a machine-learning model that can be trained to estimate unconfined compressive strength from drilling parameters instantaneously.

- Provide changes in Unconfined Compressive Strength based on drilling parameters instantaneously to avoid possible wellbore failures and drilling accidents.
- Utilize principal component analysis to analyze feature importance using available drilling data.
- Study the implementation of Random Forest regression algorithm to build a robust regression model to estimate certain geomechanical parameters (i.e., UCS, and Mechanical Specific Energy).

The initial path of this study was to implement PCA to the complete data set and use only four principal components of drilling parameters instead of seven variables (Depth, RPM, WOB, PFI, Torque, ROP, MSE) to train Random Forest regression model to predict UCS. The early exploratory analysis conducted on the data set indicated a lack of variability in UCS data as only seven different values were present even though the complete data set had 55+ million UCS observations. This lack of variability is assumed to be a unique part of the data set similar to predicted Torque values. However, as the study developed further by training the model with complete data set, a high variance error is indicated as the accuracy of UCS predictions was 98%. The high variance error was not resolved as the model was fitted numerous times with a different list of hyper-parameters. To bring a solution to this problem, a root cause of high variance error is tested through three different hypotheses. These hypotheses are as followings:

- A possible linear correlation between MSE and UCS can lead to perfect predictions.
- A loss of valuable information due to the implementation of PCA can cause overfitting.
- A lack of variability on UCS can reduce the variability of predictions, which can cause overfitting.

These hypotheses are tested to indicate the root cause of high variance error. As mentioned in Chapter 4, 1st and 2nd hypotheses are tested by changing the algorithm by removing PCA and measuring the correlation between MSE and UCS. Both 1st and 2nd hypotheses are found wrong. The 3rd hypothesis is tested by moving MSE from variables to target values as MSE would cover the variation within the data set most. The accuracy of predictions decreased by 5-10%, and the

accuracy reduction changed with respect to the amount of data used to train the model. These observations indicated that the model was robust, but the lack of variability within UCS was causing a high variance error as Multi-Output Random Forest performs similar to single output RF (Linusson 2013).

The final model is fitted with six variables (Depth, RPM, WOB, PFI, Torque, ROP) and two target values (UCS, MSE). The following conclusions are found from this study;

- Implementation of PCA on this data set will cause a valuable information loss.
- A small explained variance contribution by a variable doesn't reflect its importance in regression models as the variable might carry an important piece of information that will contribute to predictions of the model.
- A lack of variability within target values while training a regression model will cause high variance error, aka overfitting, as the number of data points fed to algorithm increase.
- This data set does not reflect the real performance of the algorithm as a high variance error is observed if the model is fitted with a single target value, even though most common tools are used to avoid overfitting.
- The model is robust enough to indicate variation change in target values as UCS predictions accuracy decreased when another target value is added to target values to introduce more variation.
- The model can be trained to estimate UCS from drilling parameters accurately if the data set with higher variations among variables is used.

The main objectives of this thesis are fulfilled as the machine learning model built by using Random Forest regression algorithm were successfully estimate UCS and MSE from drilling parameters instantaneously. Also, PCA is utilized to indicate the feature importance and the retained explained variance by each principal component to observe the possible loss of valuable information. As a result of the implementation of PCA, it is decided to use six drilling parameters directly instead of training the model with only four principal components.

CHAPTER 7

FUTURE WORK

The work completed in this study can be developed further to implement the model as an efficient data-driven solution for field applications. The suggested approaches and methods can be a part of the further development of this study as follow:

- The model built for this study can be trained with field data collected at the drilling site, and a comparison between UCS predictions from the model and UCS measurements from laboratory experiments can be made.
- Different variables affecting UCS can be introduced to the model as a part of the training data set. These variables can be porosity and elemental spectroscopy (Negara et al. 2017).
- The empirical equations derived to estimate UCS can be added to the model as a lower boundary condition (Chang et al. 2006), if the empirical equation is derived for the same formation where the data set is collected.
- The core samples retrieved from Edgar Mine by using Apache Coring Rig can be tested to measure UCS and Young's Modulus, and these measurements can be integrated into the drilling data collected. The integrated drilling parameters and UCS can be used to train the model.

REFERENCES CITED

- Al-Awad, M. N. J. 2012. Evaluation of Mohr-Coulomb Failure Criterion Using Unconfined Compressive Strength. Paper presented at the ISRM Regional Symposium - 7th Asian Rock Mechanics Symposium, Seoul, Korea. ISRM-ARMS7-2012-038.
- Al-Wardy, W. and Urdaneta, O. P. 2010. Geomechanical Modeling for Wellbore Stability during Drilling Nahr Umr Shales in a field in Petroleum Development Oman. Paper presented at the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, November 2010. SPE-138214-MS. <https://doi.org/10.2118/138214-MS>.
- AlSaihati, A., Elkatatny, S., Mahmoud, A., and Abdurraheem, A. 2021. Early Anomaly Detection Model Using Random Forest while Drilling Horizontal Wells with a Real Case Study. Paper presented at the SPE/IADC Middle East Drilling Technology Conference and Exhibition, Abu Dhabi, UAE, May 2021. SPE-202144-MS. <https://doi.org/10.2118/202144-MS>.
- American Society for Testing and Materials. 2014. ASTM D7012 Standard Test Methods for Compressive Strength and Elastic Moduli of Rock Core Specimens under Varying States of Stress and Temperatures. Technical report, American Society for Testing and Materials, (2014).
- Awad, M. and Khanna, R. 2015. Support Vector Regression. In *Efficient Learning Machines*, Chap. 4, 67–80, Berkeley, California: Apress. https://doi.org/10.1007/978-1-4302-5990-9_4.
- Barzegar, R., Sattapour, M., and Nikudel, M. R. 2016. Comparative Evaluation of Artificial Intelligence Models for Prediction of Uniaxial Compressive Strength of Travertine Rocks, Case Study: Azarshahr Area, NW Iran. *Modeling Earth Systems and Environment*, **2**(2): 1–13. <http://dx.doi.org/10.1007/s40808-016-0132-8>.
- Bourgoyne, A. T. J. and Young, F. S. J. 1974. A Multiple Regression Approach to Optimal Drilling and Abnormal Pressure Detection. *SPE J.*, **14**(04): 371–384. <http://dx.doi.org/10.2118/4238-PA>.
- Bradford, I. D. R., Fuller, J., Thompson, P. J., and Walsgrove, T. R. 1998. Benefits of Assessing the Solids Production Risk in a North Sea Reservoir Using Elastoplastic Modeling. Paper presented at the SPE/ISRM Rock Mechanics in Petroleum Engineering held in Trondheim, Norway, 8-10 July. SPE-47360-MS. <https://doi.org/10.2118/47360-MS>.
- Brehm, A., Ward, C. D., Bradford, D. W., and Darrell, E. R. 2006. Optimizing a Deepwater Subsalt Drilling Program by Evaluating Anisotropic Rock Strength Effects on Wellbore Stability and Near Wellbore Stress Effects on the Fracture Gradient. Paper presented at the IADC/SPE Drilling Conference, Miami, Florida, USA. SPE-98227-MS. <https://doi.org/10.2118/98227-MS>.

- Brook, N. 1993. The Measurement and Estimation of Basic Rock Strength. In *Comprehensive Rock Engineering*, ed. J.A. Hudson, Chap. 2, 42-66, Leeds, UK: Pergamon Press. <https://doi.org/10.1016/B978-0-08-042066-0.50009-4>.
- Ceryan, N., Okkan, U., and Kesimal, A. 2013. Prediction of Unconfined Compressive Strength of Carbonate Rocks Using Artificial Neural Networks. *Environmental Earth Sciences*, **68**: 807–819. <http://dx.doi.org/10.1007/s12665-012-1783-z>.
- Chang, C., Zoback, M., and Khaksar, A. 2006. Empirical Relations between Rock Strength and Physical Properties in Sedimentary Rocks. *J Pet Technol*, **51**: 223–237. <http://dx.doi.org/10.1016/j.petrol.2006.01.003>.
- Chen, C., Han, X., Yang, M., Zhang, W., Wang, X., and Peng, D. 2019. A New Artificial Intelligence Recognition Method of Dominant Channel Based on Principal Component Analysis. Paper presented at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Bali, Indonesia, October 2019. SPE-196295-MS. <https://doi.org/10.2118/196295-MS>.
- Combs, G. D. 1968. Prediction of Pore Pressure from Penetration Rate. Paper presented at the Fall Meeting of the Society of Petroleum Engineers of AIME, Houston, September 29 - October 2. SPE-2162-MS. <https://doi.org/10.2118/2162-MS>.
- Cunningham, R. and Eenink, J. G. 1959. Laboratory Study of Effects of Overburden, Formation and Mud Column Pressures on Drilling Rate of Permeable Formations. *Transaction of the Society of Petroleum Engineers*, **217**: 9–17. <https://doi.org/10.2118/1094-G>.
- de lima, R. P. and Marfurt, K. J. 2018. Principal component analysis and K-means analysis of airborne gamma-ray spectrometry surveys. Paper presented at the 2018 SEG International Exposition and Annual Meeting, Anaheim, California, USA, October 2018. SEG-2018-2996506. <https://doi.org/10.1190/segam2018-2996506.1>.
- Elghonimy, R. and Sonnenberg, S. 2021. A Principal Component Analysis Approach to Understanding Relationships Between Elemental Geochemistry Data and Deposition, Niobrara Formation, Denver Basin, CO. Paper presented at the SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA, July 2021. URTEC-2021-5440-MS. <https://doi.org/10.15530/urtec-2021-5440>.
- Fjær, E., Holt, R. M., and Horsrud, P. 1992. Mechanical Properties from Field Data. In *Petroleum Related Rock Mechanics*, ed. E. Fjær, R.M. Holt, P. Horsrud, A.M. Raaen, R. Risnes, Chap. 8, 209-236, Amsterdam, The Netherlands: Elsevier. [https://doi.org/10.1016/S0376-7361\(09\)70194-3](https://doi.org/10.1016/S0376-7361(09)70194-3).
- Ford, C. 2015. Is R-squared Useless? 6th September. <https://data.library.virginia.edu/is-r-squareduseless/> (accessed 23 September 2020).
- Geron, A. 2019. The Machine Learning Landscape. In *Hands-on Machine Learning with Scikit-learn, Keras, and TensorFlow: Concept*, ed. Aurelien G., Chap. 1, 3–34, Sebastopol, California: O'Reily Media Inc.

- Gstalder, S. and Raynal, J. 1966. Measurement of Some Mechanical Properties of Rocks and Their Relationship to Rock Drillability. *J Pet Technol*, **18**(08): 991–996. <http://dx.doi.org/10.2118/1463-PA>.
- Guo, H. K., Marfurt, K. J., and Liu, J. 2009. Principal Component Spectral Analysis. *Geophysics*, **74**(04): 35–43. <http://dx.doi.org/10.1190/1.3119264>.
- Gupta, S., Saputelli, L., and Nikolaou, M. 2016. Applying Big Data Analytics to Detect, Diagnose, and Prevent Impending Failures in Electric Submersible Pumps. Paper presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE, September 2016. SPE-181510-MS. <https://doi.org/SPE-181510-MS>.
- Hassanpour, J., Rostami, J., and Zhao, J. 2011. A New Hard Rock TBM Performance Prediction Model for Project Planning. *Tunnelling and Underground Space Technology*, **26**(05): 595–603. <http://dx.doi.org/10.1016/j.tust.2011.04.004>.
- Hegde, C., Wallace, S., and Gray, K. 2015. Using Trees, Bagging, and Random Forests to Predict Rate of Penetration During Drilling. Paper presented at the SPE Middle East Intelligent Oil and Gas Conference and Exhibition, Abu Dhabi, UAE, September 2015. SPE-176792-MS. <https://doi.org/10.2118/176792-MS>.
- Horshud, P. 2001. Estimating Mechanical Properties of Shale from Empirical Correlations. *SPE Drilling & Completion*, **16**(02): 68–73. <http://dx.doi.org/10.2118/56017-PA>.
- Iferobia, C. C., Ahmad, M., Salim, A. M., Sambo, C., and Ifechukwu, H. M. 2020. Acoustic Data Driven Application of Principal Component Multivariate Regression Analysis in the Development of Unconfined Compressive Strength Prediction Models for Shale Gas Reservoirs. SPE-201287-MS. Paper presented at the SPE Annual Technical Conference and Exhibition, Virtual, October 2020. SPE-201287-MS. <https://doi.org/10.2118/201287-MS>.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. Linear Regression. In *An Introduction to Statistical Learning*, ed. Gareth J., Daniela W., Trevor H., Robert H., Chap. 3, 59–126, New York, NY: Springer. https://doi.org/10.1007/978-1-4614-7138-7_3.
- Jorden, J. R. and Shirley, O. J. 1966. Application of Drilling Performance Data to Overpressure Detection. *JPT Pet Technol*, **18**(11): 1387–1394. <http://dx.doi.org/10.2118/1407-PA>.
- Joshi, D. R. 2021. Real-time Characterization of Water-Bearing Lunar Regolith Drilling on The Moon. MS thesis, Colorado School of Mines, Golden, Colorado (May 2021).
- Klimentos, T. 2005. Optimizing Drilling Performance by Wellbore Stability and Pore-Pressure Evaluation in Deepwater Exploration.” Paper presented at the International Petroleum Technology Conference, Doha, Qatar, November 2005. doi: <https://doi.org/10.2523/IPTC-10933-MS>.
- Kong, X., Hu, C., and Duan, Z. 2017. Introduction. In *Principal Component Analysis Networks and Algorithms*, ed. X. Kong, C. Hu, Z. Duan, Chap. 1, 1-16, Singapore: Springer. https://doi.org/10.1007/978-981-10-2915-8_1.

- Lal, M. 1999. Shale Stability: Drilling Fluid Interaction and Shale Strength. Paper presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, Jakarta, Indonesia, April 1999. SPE-54356-MS. <https://doi.org/10.2118/54356-MS>.
- Lashkaripour, G. R. and Dusseault, M. B. 1993. A Statistical Study on Shale Properties: Relationships among Principal Shale Properties. In *Probabilistic Methods in Geotechnical Engineering*, ed. K.S. Li, S-C.R. Lo, Chap. 4, 191-197, Udine, Italy: CRC Press. <https://doi.org/10.1201/9781003077749>.
- Linusson, H. 2013. Multi Output Random Forests. Master thesis, University of Borås, Borås, Sweden (2013).
- Løken, E. A., Løkkevik, J., and Sui, D. 2020. Data-driven Approaches Tests on a Laboratory Drilling System. *Journal of Petroleum Exploration and Production Technology*, **10**: 3043–3055. <https://doi.org/10.1007/s13202-020-00870-z>.
- Majidi, R., Albertin, M., and Nigel, L. 2017. Pore-pressure Estimation by Use of Mechanical Specific Energy and Drilling Efficiency. *SPE Drilling & Completion*, **32**(02): 097–104. <http://dx.doi.org/10.2118/178842-PA>.
- McNally, G. H. 1987. Estimation of Coal Measures Rock Strength Using Sonic and Neutron Logs. *Geoexploration*, **24**(04): 381–395. [http://dx.doi.org/10.1016/0016-7142\(87\)90008-1](http://dx.doi.org/10.1016/0016-7142(87)90008-1).
- Meulenkamp, F. and Alvarez, G. M. 1999. Application of Neural Networks for The Prediction of The Unconfined Compressive Strength (ucs) from Equotip Hardness. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, **36**(01): 29–39. [http://dx.doi.org/10.1016/S0148-9062\(98\)00173-9](http://dx.doi.org/10.1016/S0148-9062(98)00173-9).
- Moos, D., Zoback, M. D., and Bailey, L. 1999. Feasibility Study of the Stability of Openhole Multilaterals, Cook Inlet, Alaska. Paper presented at the SPE Mid-Continent Operations Symposium held in Oklahoma City, Oklahoma, 28-31 March. SPE-52186-MS. <https://doi.org/10.2118/52186-MS>.
- Nabaei, M., Shahbazi, K., and Shadravan, A. 2010. Uncertainty Analysis in Unconfined Rock Compressive Strength Prediction. Paper presented at the SPE Deep Gas Conference and Exhibition, Manama, Bahrain, January 2010. <https://doi.org/10.2118/131719-MS>.
- Nasir, E. and Rickabaugh, C. 2018. Optimizing Drilling Parameters Using a Random Forests ROP Model in the Permian Basin. Paper presented at the SPE Liquids-Rich Basins Conference - North America, Midland, Texas, USA, September 2018. SPE-191796-MS. <https://doi.org/10.2118/191796-MS>.
- Negara, A., Ali, S., Aldhamen, A., Kesserwan, H., and Guodong, J. 2017. Unconfined Compressive Strength Prediction from Petrophysical Properties and Elemental Spectroscopy Using Support-Vector Regression. Paper presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, April 2017. SPE-188077-MS. <https://doi.org/10.2118/188077-MS>.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**: 2825–2830.
- Rehm, B. and McClendon, R. 1971. Measurement of Formation Pressure from Drilling Data. Presented at the Fall Meeting of the Society of Petroleum Engineers, New Orleans, 3–6 October. SPE-3601-MS. <https://doi.org/10.2118/3601-MS>.
- Shalizi, C. 2015. Lecture 10: F-Tests, R^2 , and Other Distractions. <http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/10/lecture-10.pdf> (accessed 16 September 2020).
- Song, M. and Zhou, X. 2019. A Casing Damage Prediction Method Based on Principal Component Analysis and Gradient Boosting Decision Tree Algorithm. Paper presented at the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, March 2019. SPE-194956-MS. <https://doi.org/10.2118/194956-MS>.
- Spaar, J. R., Ledgerwood, L. W., Goodman, H., Graff, R. L., and Moo, T. J. 1995. Formation Compressive Strength Estimates for Predicting Drillability and PDC Bit Selection. Paper presented at the SPE/IADC Drilling Conference, Amsterdam, Netherlands, February 1995. <https://doi.org/10.2118/29397-MS>.
- Swalin, A. 2018. Choosing the Right Metric for Evaluating Machine Learning Models. 5th September. <https://medium.com/usf-msds/choosing-the-right-metric-for-machinelearning-models-part-1-a99d7d7414e4> (accessed 5 December 2020).
- Tarkoy, P. J. and Marconi, M. 1991. Difficult Rock Comminution and Associated Geological Conditions. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, **29**(04): 262. [http://dx.doi.org/10.1016/0148-9062\(92\)90959-4](http://dx.doi.org/10.1016/0148-9062(92)90959-4).
- Teale, R. 1965. The Concept of Specific Energy in Rock Drilling. *International Journal of Rock Mining Science*, **2**(01): 57–73. [http://dx.doi.org/10.1016/0148-9062\(65\)90022-7](http://dx.doi.org/10.1016/0148-9062(65)90022-7).
- Van Rossum, G. and Drake, F. L. 2009. *Python 3 Reference Manual*. CreateSpace.
- Vernik, L., Bruno, M., and Bovberg, C. 1993. Empirical Relations between Compressive Strength and Porosity of Siliciclastic Rocks. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, **30**(07): 677–680. [http://dx.doi.org/0148-9062\(93\)90004-W](http://dx.doi.org/0148-9062(93)90004-W).
- Yurdakul, M., Ceylan, H., and Hurriyet, A. 2011. "A Predictive Model For Uniaxial Compressive Strength of Carbonate Rocks From Schmidt Hardness." Paper presented at the 45th U.S. Rock Mechanics / Geomechanics Symposium, San Francisco, California, June 2011.
- Zhang, J. J. 2019. Rock Strengths and Rock Failure Criteria. In *Applied Petroleum Geomechanics*, ed. J.J. Zhang, Chap. 2, 29-83, Oxford, UK: Gulf Professional Publishing. <https://doi.org/10.1016/B978-0-12-814814-3.00002-2>.

APPENDIX A
SUPPLEMENTAL FILE

Permission to include Figure 2.1 from The Applied Petroleum Geomechanics, Vol. I, Zhang,J,J, Chapter III, 86, Copyright (2020) has been obtained from Elsevier; the copyright permission statement may be found in Supplemental File Elsevierpermission.pdf.

Permissions to include Figures reproduced from other published sources in included in Appendix B.


APPENDIX B
COPYRIGHT PERMISSIONS


Permissions to include previously published material in this thesis is below.

1) The Figure 4.3 is reproduced from Efficient Learning Machines, Chapter IV, which is open access to public.

2) Permission from D.R. Joshi regarding the reproduction of materials from his thesis, Joshi (2021), is expressed below.

[External] Re: Publication Copyright Permission

 Deep Joshi <deep.joshi@corva.ai>
To: Muhammed Kaya

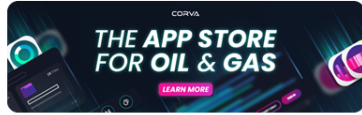
 If there are problems with how this message is displayed, click here to view it in a web browser.



Dear Mr. Kaya,

You have my permission to use these figures from the work.

Good luck.

Deep Joshi Jr. Research Engineer
Mobile: 346-285-1786 24/7 Ops: 281-742-9370





 Reply  Reply All  Forward  

Mon 4/18/2022 3:07 PM

3) Permission from Cassandra Furtado on behalf of O'Reilly Media Inc. regarding Figure 4.2, and Figure 4.4 is expressed below.

[External] Re: Urgent Publication Permission Needed






 Cassandra Furtado <cfurtado@oreilly.com>
To: Muhammed Kaya
Cc: Teri Finn

 You replied to this message on 3/31/2022 1:39 PM.
Click here to download pictures. To help protect your privacy, Outlook prevented automatic download of some pictures in this message.

Hello Muhammed,

Thank you for reaching out to us. O'Reilly Media Inc. is pleased to grant you permission to use those two figures in your thesis, as described below, free of charge. Please include the following credit where applicable: "from *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, by Aurélien Géron. Copyright © 2019 Kiwisoft S.A.S. Published by O'Reilly Media, Inc. Used with permission."

Best wishes,
Cassandra

 Reply  Reply All  Forward  

Thu 3/31/2022 1:22 PM