KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

KUMASI, GHANA

**COLLEGE OF ENGINEERING**

DEPARTMENT OF PETROLEUM ENGINEERING

**PROJECT REPORT**

TOPIC:PREDICTION OF ROCK STRENGTH FROM GAMMA RAY AND RATE OF PENETRATION DATA USING RANDOM FOREST, LASSO REGRESSION AND DECISION TREE MACHINE LEARNING ALGORITHMS.

**PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE**

**REQUIREMENTS FOR THE BSC (ENG.) DEGREE**

|  |  |
| --- | --- |
| Name | Index No |
| Mbil Abdul Rahman | 2943220 |
| Appiah Nhyiraba Paa Kwesi | 2933320 |
| Obeng Clinton | 2945420 |
| Appianing Jude Nana Ahenkorah | 2933520 |
| Quaidoo Joshua | 2947320 |
| Mensah Daniel | 2943420 |
| Acquaah Aba Akyeneba | 2926520 |
| Kwei Richmond Mensah | 2942520 |

SUPERVSOR: STEPHEN ADJEI, PHD

**AUGUST,2024**

# DECLARATION

We affirm that the project titled “Prediction of Rock Strength from Gamma Ray and Rate Of Penetration Data using Random Forest, Lasso Regression And Decision Tree Machine Learning Algorithms” presented to the Department of Petroleum Engineering at Kwame Nkrumah University of Science and Technology is an original work conducted by us under the guidance of Dr. Stephen Adjei, a faculty member of the Petroleum Programme within the College of Engineering. This project serves as part of the requirements for the completion of our Bachelor of Science Degree in Petroleum Engineering. The findings and content included in this report have not been previously submitted to any other academic institution for a degree or diploma.

|  |  |
| --- | --- |
| MBIL ABDUL RAHMAN | ………………………………………….. |
| APPIAH NHYIRABA PAA KWESI | ………………………………………….. |
| QUAIDOO JOSHUA | ………………………………………….. |
| KWEI RICHMOND MENSAH | ………………………………………….. |
| APPIANING JUDE NANA AHENKORAH | ………………………………………….. |
| OBENG CLINTON | ………………………………………….. |
| ACQUAAH ABA AKYENEBA | ………………………………………….. |
| MENSAH DANIEL | ………………………………………….. |

# SUPERVISOR DECLARATION

It is certified that this project has been prepared and submitted under my supervision.

DATE ………………………………

SUPERVISOR’S SIGNATURE ………………………………

Dr. Stephen Adjei

# DEDICATION

This thesis is first and foremost dedicated to the Almighty God and our supervisor Dr. Stephen Adjei. We also dedicate the work to our lovely family.

# ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Stephen Adjei for his invaluable guidance and support throughout the completion of this project. His expertise and mentorship have been instrumental in shaping our research and academic growth.

# ABSTRACT

Given the challenges of drilling through varying subsurface formations, developing advanced techniques to quickly and effectively predict rock strength is key to maintaining the efficiency and safety of oil and gas operations since conventional methods are often time-consuming and destructive to the core samples. In recent years, several studies have been conducted to estimate rock strength using non-destructive tests. However, proper analysis are often not conducted on the machine learning models to understand their characteristics and reliability before they are implemented. This project aims to address this gap by thoroughly evaluating various machine learning models to ensure their effectiveness in predicting rock strength accurately.

The data obtained consisted of Rate of Penetration and Gamma Ray measurements, which were preprocessed and used to develop machine learning models such as LASSO Regression, Decision Tree, and Random Forest. Python libraries, including NumPy, Pandas, Scikit-learn, Matplotlib, and Seaborn, were employed to implement these models, enabling effective feature analysis and prediction of rock strength.

The results of this study revealed that the models performed poorly due to large prediction errors, suggesting potential overfitting. The significant errors indicated that the current models may not be capturing all relevant features influencing rock strength, highlighting the need for further refinement and feature engineering to improve predictive accuracy.

TABLE OF CONTENTS

[DECLARATION I](#_Toc175862438)

[SUPERVISOR DECLARATION II](#_Toc175862439)

[DEDICATION III](#_Toc175862440)

[ACKNOWLEDGEMENT IV](#_Toc175862441)

[ABSTRACT V](#_Toc175862442)

[LIST OF EQUATIONS XI](#_Toc175862443)

[NOMENCLATURE XII](#_Toc175862444)

[CHAPTER 1 1](#_Toc175862445)

[1.1 BACKGROUND 1](#_Toc175862446)

[1.2 PROBLEM STATEMENT 4](#_Toc175862447)

[1.3 AIMS AND OBJECTIVES 5](#_Toc175862448)

[1.3.1 AIM 5](#_Toc175862449)

[1.3.2 OBJECTIVES 5](#_Toc175862450)

[1.4 RESEARCH JUSTIFICATION 5](#_Toc175862451)

[1.5 LIMITATIONS 6](#_Toc175862452)

[1.6 STRUCTURAL OUTLINE 6](#_Toc175862453)

[CHAPTER TWO 8](#_Toc175862454)

[LITERATURE REVIEW 8](#_Toc175862455)

[CHAPTER 3 20](#_Toc175862456)

[METHODOLOGY 20](#_Toc175862457)

[3.1 OVERVIEW 20](#_Toc175862458)

[3.2 DATA PREPROCESSING 22](#_Toc175862459)

[3.2.1 DATA CLEANING 22](#_Toc175862460)

[3.2.2 FEATURE SCALING 26](#_Toc175862461)

[3.2.3 CORRELATION MATRIX AND HEAT MAP 27](#_Toc175862462)

[3.3 EXPLORATORY DATA ANALYSIS 28](#_Toc175862463)

[3.3.1 STATISTICAL MEASURES 28](#_Toc175862464)

[3.3.2 DATA VISUALIZATION 29](#_Toc175862465)

[3.3.3 DATA SPLITTING 29](#_Toc175862466)

[3.4 EVALUATION METRICS 30](#_Toc175862467)

[3.4.1 MEAN ABSOLUTE ERROR ( MAE) 30](#_Toc175862468)

[3.4.2 MEAN SQUARE ERROR 31](#_Toc175862469)

[3.4.3 COEFFICIENT OF DETERMINATION 31](#_Toc175862470)

[CHAPTER 4 32](#_Toc175862471)

[RESULTS AND DISCUSSION 32](#_Toc175862472)

[4.1 DECISION TREE MODEL 32](#_Toc175862473)

[4.2 RANDOM FOREST MODEL 35](#_Toc175862474)

[4.3 LASSO REGRESSION 38](#_Toc175862475)

[CHAPTER 5 41](#_Toc175862476)

[CONCLUSION AND RECOMMENDATIONS 41](#_Toc175862477)

[5.1 CONCLUSION 41](#_Toc175862478)

[5.2 RECOMMENDATION 42](#_Toc175862479)

[REFERENCES 44](#_Toc175862480)

**LIST OF FIGURES**

[FIGURE 1. LABORATORY TESTS FOR ROCK STRENGTH PREDICTIONS 2](#_Toc175867973)

[FIGURE 2. MACHINE LEARNING ALGORITHMS 3](#_Toc175867974)

[FIGURE 3. MINIATURE DISC CUTTER USED FOR THIS PROJECT (RULER IN INCHES) 9](#_Toc175867975)

[FIGURE 4. LINEAR CUTTING DEVICE DEVELOPED TO RUN THE SCRATCH TESTS 9](#_Toc175867976)

[FIGURE 5. PICTURES OF ROCK SAMPLES PREPARED FOR CUTTING TEST 10](#_Toc175867977)

[FIGURE 6. TYPICAL ARCHITECTURE OF FIS MODEL 11](#_Toc175867978)

[FIGURE 7. SCHEMATIC SHOWING THE ANN MODEL 12](#_Toc175867979)

[FIGURE 8. SCHEMATIC OF THE FIVE-LAYER ARCHITECTURE OF THE ANFIS 12](#_Toc175867980)

[FIGURE 9. WELL LOGS 14](#_Toc175867981)

[FIGURE 10. SUPPORT VECTOR MACHINE 15](#_Toc175867982)

[FIGURE 11. RANDOM FOREST ALGORITHM 16](#_Toc175867983)

[FIGURE 12. GENERAL ARCHITECTURE OF XGBOOST 17](#_Toc175867984)

[FIGURE 13. SUPPORT VECTOR REGRESSION 17](#_Toc175867985)

[FIGURE 14. BOX PLOT OF ACTUAL AND PREDICTED UCS VALUES 19](#_Toc175867986)

[FIGURE 15 - GENERAL WORKFLOW 21](#_Toc175867987)

[FIGURE 16 - TREND IN DATA 22](#_Toc175867988)

[FIGURE 17 - NULL VALUES IN DATAFRAME 23](#_Toc175867989)

[FIGURE 18 - LINEAR INTERPOLATION OF DEPTH COLUMN 24](#_Toc175867990)

[FIGURE 19 - INTERPOLATION OF ROP COLUMN 24](#_Toc175867991)

[FIGURE 20 - INTERPOLATION OF GR COLUMN 24](#_Toc175867992)

[FIGURE 21 - INTERPOLATION OF UCS COLUMN 24](#_Toc175867993)

[FIGURE 22 - VISUALIZING CLEAN DATA 25](#_Toc175867994)

[FIGURE 23 - CORRELATION MATRIX AND HEATMAP 27](#_Toc175867995)

[FIGURE 24 - DATA SPLITTING ARCHITECTURE 30](#_Toc175867996)

[FIGURE 26 - TRAIN DATA PLOT FOR DECISION TREE MODEL 33](#_Toc175867997)

[FIGURE 27 – TEST DATA PLOT FOR DECISION TREE MODEL 34](#_Toc175867998)

[FIGURE 29 - TRAIN DATA PLOT FOR RANDOM FOREST 36](#_Toc175867999)

[FIGURE 30 - TEST DATA PLOT FOR RANDOM FOREST 37](#_Toc175868000)

[FIGURE 32 - TRAIN DATA PLOT FOR LASSO REGRESSION 39](#_Toc175868001)

[FIGURE 33 - TEST DATA PLOT FOR LASSO REGRESSION 40](#_Toc175868002)

**LIST OF TABLES**

[TABLE 1 - EVALUATION METRICS FOR DECISION TREE MODEL 43](#_Toc175689240)

[TABLE 2 - EVALUATION METRICS FOR RANDOM FOREST MODEL 46](#_Toc175689241)

[TABLE 3 – EVALUATION METRICS FOR LASSO REGRESSION MODEL 49](#_Toc175689242)

# LIST OF EQUATIONS

EQUATION [(1) 2](#_Toc175689361)

EQUATION [(2) 2](#_Toc175689362)

EQUATION [(3) 41](#_Toc175689363)

EQUATION [(4) 41](#_Toc175689364)

EQUATION [(5) 41](#_Toc175689365)

# NOMENCLATURE

* ANN – Artificial Neural Network
* FIS – Fuzzy Inference System
* ANFIS – Adaptive Neuro – Fuzzy Inference System
* TCS – Tri-axial Compressive Strength
* UCS – Uniaxial Compressive Strength
* 𝑅𝑛 - Schmidt Rebound Number
* 𝑉𝑝 - P-wave velocity
* MLP – Multilayer Perception
* GEP – Genetic Expression Programming
* GP – Genetic Programming
* LS-SVM – Least Squares Support Vector Machine

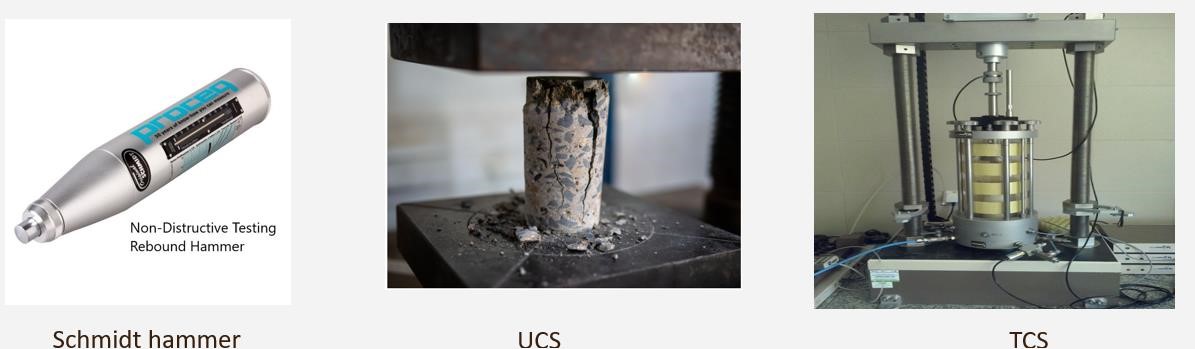
# CHAPTER 1

## 1.1 BACKGROUND

Rock strength refers to the ability of a rock material to withstand the applied forces without breaking or deforming (Goel, 2012). Rock strength is a fundamental property that plays a vital role in various industries such as geology, civil engineering, and mining. This property is crucial in understanding the behavior of rocks under different conditions and stresses. Rock strength is influenced by factors such as mineral composition, grain size, porosity, and the presence of structural features like faults and fractures (Blowers, 2003). There are different types of rock strength that petroleum engineers consider when analyzing reservoir formation. Some of the key types of rock strength include compressive strength, tensile strength, and shear strength. Each type of strength provides valuable information about how rocks respond to various forces and stresses, which is crucial for designing drilling programs, wellbore stability, assessing reservoir quality, and optimizing production in the oil and gas industries.

Several methods are employed to predict rock strength:

1. Laboratory Testing: This involves conducting various laboratory tests on rock core samples collected from the site. Common tests include uniaxial compressive strength (UCS), Triaxial Compressive Strength (TCS), Schmidt Hammer. These tests provide direct measurements of the rock's strength properties under controlled conditions.



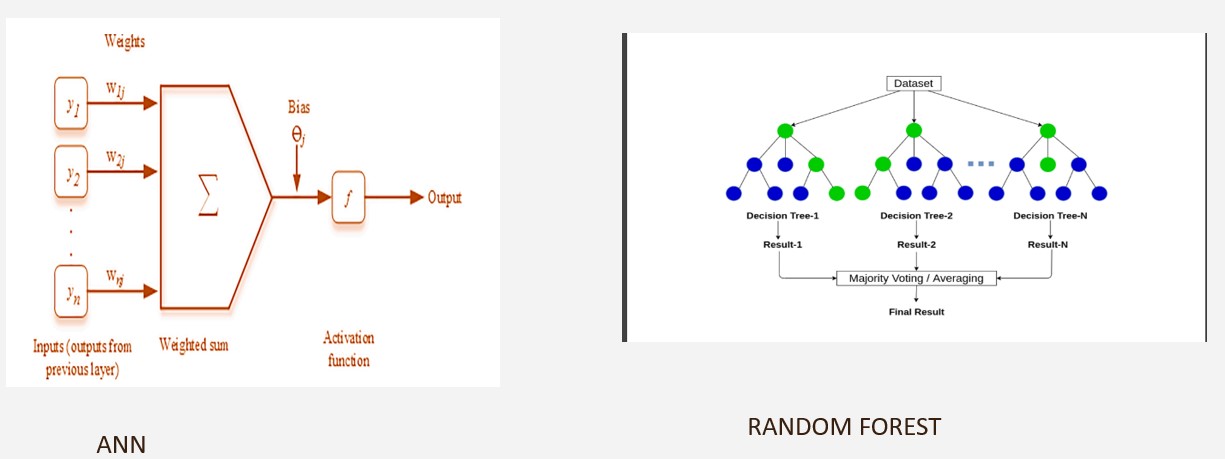
**Figure 1. Laboratory tests for Rock Strength Predictions**

1. Empirical Methods: Empirical correlations utilize observed relationships between rock properties and strength parameters derived from field and laboratory data. An example is the Mohr-Coulomb. These methods are relatively quick and inexpensive but may lack accuracy for specific geological conditions *(Yuan et al., 2024).*

|  |  |
| --- | --- |
|  | (1) |

|  |  |
| --- | --- |
|  | (2) |

1. Machine Learning Algorithms: They are used to predict rock strength by analyzing various data inputs related to the composition, structure, and properties of rocks ( *et al.*, 2024). These algorithms can be trained on a dataset that includes information such as mineralogy, porosity, grain size, and other geological characteristics of different rock samples along with their corresponding strength values. Examples of models used are ANN, Random Forest.



**Figure 2. Machine Learning Algorithms**

## 1.2 PROBLEM STATEMENT

The accurate prediction of rock strength plays a crucial role in various geotechnical and petroleum engineering applications, including wellbore stability analysis, hydraulic fracturing design, and reservoir characterization *(Dusim, 2020*). The conventional ways of rock strength prediction are destructive, costly and time consuming. Also, the complexities of geological and field data can present challenges in accurately interpreting results. Furthermore, the reliance on empirical models can lead to errors due to many assumptions made during analysis. Therefore, the machine learning algorithms would help predict rock strength rapidly, efficiently and effectively. In recent years, machine learning algorithms have emerged as promising tools for predicting rock properties using readily available drilling data, such as gamma ray measurements and rate of penetration (ROP). Random Forest, Lasso Regression, and Decision Tree models are among the widely used algorithms for such predictive tasks due to their simplicity, interpretability, and ability to handle complex data relationships *(Manoharan, 2024*). Despite the growing interest in using machine learning to estimate rock strength from ROP and gamma ray data, a comprehensive comparison of various methods is still needed to understand their effectiveness and identify the best approach for accurately estimating rock strength across different geological contexts.

## 1.3 AIMS AND OBJECTIVES

### 1.3.1 AIM

The aim of this work is to compare and harness the strengths of LASSO Regression, Random Forest and Decision Tree Machine Learning Algorithms for predicting rock strength in order to provide more accurate and interpretable reliable predictions.

### 1.3.2 OBJECTIVES

To use evaluation metrics like Co-efficient of determination (𝑅2), Root Mean Square Error (RMSE), Mean Square Error (MAE), Mean Square Error (MSE) to evaluate the performance of models by comparing predicted values to actual values in a structured format.

## 1.4 RESEARCH JUSTIFICATION

Comprehensive knowledge and analysis of in situ rock strength and geo-mechanical characteristics of rocks are fundamental in hydrocarbon and mineral exploration stage to maximize wellbore performance, maintain wellbore stability, and optimize hydraulic fracturing process (*Weng et al.,2024*). This study is to use matrix evaluators from Machine Learning Algorithms (Random Forest, LASSO Regression and Decision Tree) to predict rock strength from gamma ray and rate of penetration data due to the high cost of laboratory-based measurements of rock strength and too many assumptions made from the empirical models.

## 1.5 LIMITATIONS

The study aims to develop a more convenient method for predicting rock strength using machine learning algorithms, leveraging gamma ray and rate of penetration data to ensure wellbore stability and inform drill bit selection and procurement decisions. However, the accuracy of these predictions may be impacted by various external factors, including weathering, stress conditions, and mineral composition, which may not be fully captured by the input data. Furthermore, additional external factors such as sensor failure can compromise data integrity, affecting the reliability of the predicted rock strength parameters. These limitations highlight the need for careful consideration of potential external influences and data quality to ensure robust and accurate predictions.

## 1.6 STRUCTURAL OUTLINE

This study is structured into five chapters. The first chapter is an introduction of the thesis paper which establishes the groundwork by discussing the background, aims, objectives, justification for the research, significance, and limitations associated with predicting rock strength through machine learning algorithms. The second chapter illustrates essential concepts and keywords relevant to rock strength prediction, including an overview of algorithms such as Random Forest, Lasso Regression, and Decision Trees, thus creating a robust understanding of the topic. The third chapter describes the machine learning workflow which involved the dataset used, the data preprocessing steps taken, feature selection processes, and the experimental design utilized for training and testing the models, thereby outlining the study's procedural framework. Subsequently, the fourth chapter presents the findings from the machine learning models, compares their performance, and discusses these results, stating the robustness and weakness of the chosen models. Finally, the fifth chapter combines the key findings, wraps up the study, and offers suggestions for future research and practical applications, thereby highlighting the importance of this research.

# CHAPTER TWO

# LITERATURE REVIEW

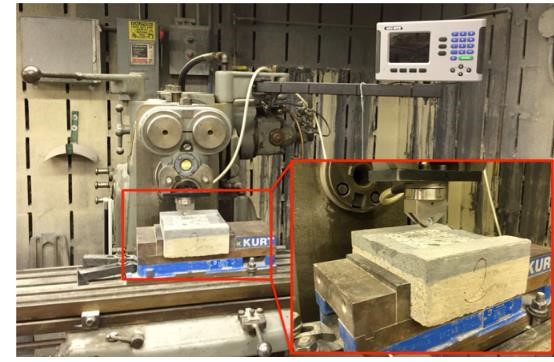
The accurate determination of rock parameters like Uniaxial Compressive Strength (UCS) is essential for comprehensive rock mechanics studies. However, laboratory experiments to obtain these parameters can be both time-consuming and costly. Consequently, researchers have been exploring more efficient methods for estimating these parameters. Traditional regression tools, though widely used, often fall short in delivering accurate predictions across diverse datasets. To overcome these limitations, recent studies have increasingly turned to machine learning (ML) methods, which have shown promising results in creating models that cover a broader range of data. Techniques such as Extreme Gradient Boosting (XGBoost), Support Vector Regression, Ridge Regression, Decision Tree, Random Forest, and LASSO Regression have proven to be particularly effective in predicting UCS (Li and Zhao, 2021)*.* The prediction of UCS is vital in various engineering and geological applications. Traditionally, multilinear regression models have been employed for this purpose. However, with the advancements in machine learning techniques, there is a growing interest in assessing their efficacy for UCS prediction. This literature review aims to explore the applicability of various traditional ways and distinct ML models in predicting the strength rocks (Li, Li and Zhang, 2022)*.*

In a study conducted by A. Naeimipour, J. Rostami, I.S Buyuksagis, and O. Frough, the authors aimed to estimate the strength of rocks by utilizing a scratch test method involving a miniature disc cutter on rock cores or inside boreholes. This method offers significant advantages over conventional approaches, which are often time-consuming, expensive, and require specialized equipment. The study focused on predicting Uniaxial Compressive Strength (UCS) and Brazilian

Tensile Strength (BTS) for sedimentary, metamorphic, and igneous rocks using the scratch test method.

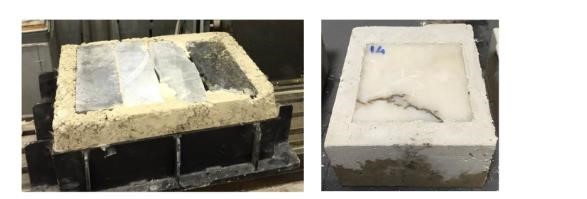


**Figure 3. Miniature disc cutter used for this project (ruler in inches)**



**Figure 4. Linear cutting device developed to run the scratch tests**

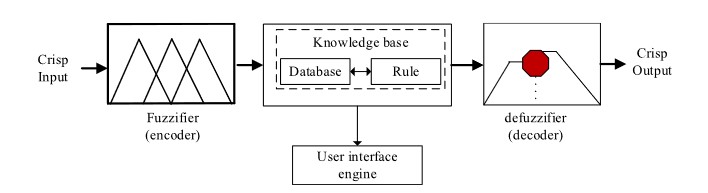
The study involved experiments with 27 different rock samples, each with varying compositions, grain size distributions, and mineral constituents. During the experiments, the average normal forces exerted on each rock sample by the miniature disc cutter were recorded. For each test, variables such as rock type, average normal force, and surface description of the samples were measured and documented (Li, Li and Zhang, 2022).



**Figure 5. Pictures of rock samples prepared for cutting test**

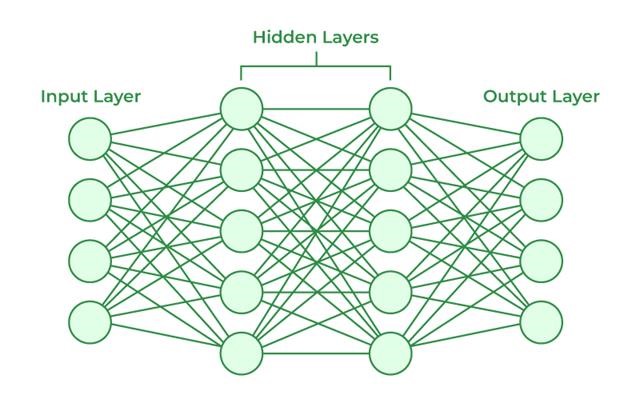
The correlations between these parameters and the strength properties of the rock samples, specifically UCS and BTS, were then analyzed. To establish the relationship between the average normal force and the depth of penetration, various regression functions (linear, power, logarithmic, quadratic, and exponential) were applied. A MATLAB code was developed for the data analysis, and R² values were observed for all the rock samples studied(Naeimipour et al., 2018).

In a study by Maryam Parsajoo and Daniel Jahed Armaghani, the tensile strength of rock materials, crucial for slope stability, tunnelling, and mining, was explored using non-destructive testing (NDT) methods. Tensile strength, the ability of a material to withstand forces such as pulling or stretching, is traditionally measured in laboratories through direct sample failure, which is both destructive and time-consuming. As an alternative, NDT methods assess the properties of the rock without causing any damage to the specimen. This review focuses on the recent advancements in using NDTs to predict the Brazilian Tensile Strength (BTS) of rocks. The evaluation of these methods involved four statistical measures: the coefficient of determination (R²), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Scatter Index (SI). The input data for this study were obtained from three nondestructive tests performed on 127 granitic rock samples: the Schmidt hammer, p-wave velocity, and density measurements. The results from these tests were used as input parameters for predictive models. Three different models were employed in the prediction of BTS. The first model, the Fuzzy Inference System (FIS), consists of four components: a fuzzifier, knowledge base, fuzzy output engine, and defuzzifier. The fuzzifier transforms crisp inputs into fuzzy values using Membership Functions (MFs) stored in the database. The knowledge base contains a collection of fuzzy rules formed using MFs in the format of If–Then rules. The fuzzy output engine applies these rules to generate output, typically using minimization and product as inference operators. The defuzzifier then converts the fuzzy output back into a crisp value.



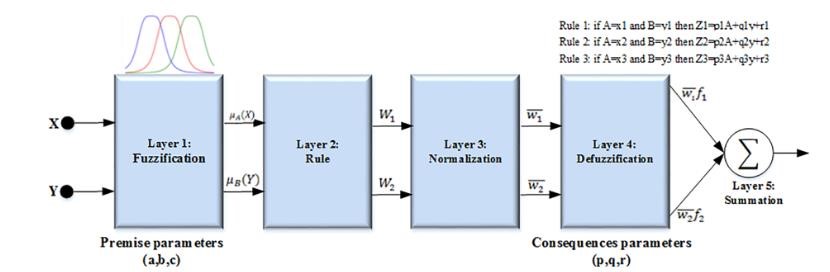
**Figure 6. Typical architecture of FIS model**

The second model used was an Artificial Neural Network (ANN), where the sensory input is represented by neuronal activation. In an ANN, different layers of features are extracted, with each layer's features being combinations of those from the previous layer. A standard ANN consists of an input layer that receives data and passes it to the next layer, where it is processed by an activation function before being sent to subsequent layers.



**Figure 7. Schematic showing the ANN model**

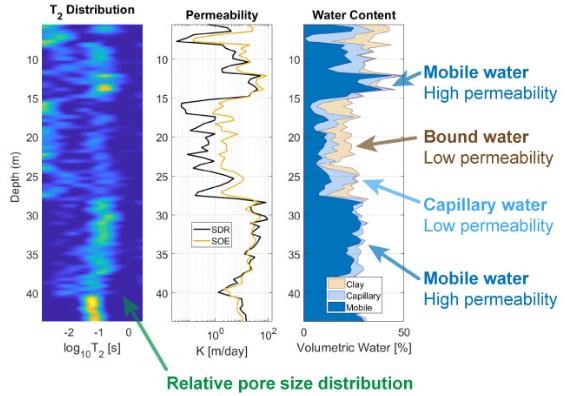
The third model, the Adaptive Neuro-Fuzzy Inference System (ANFIS), is designed to map multiple inputs to a single output for prediction tasks. This mapping is achieved by defining Membership Functions for each input variable. During training, the parameters of the premise and consequence layers are optimized. ANFIS generally consists of five layers: fuzzification, rule, normalization, defuzzification, and summation layers.



**Figure 8. Schematic of the five-layer architecture of the ANFIS**

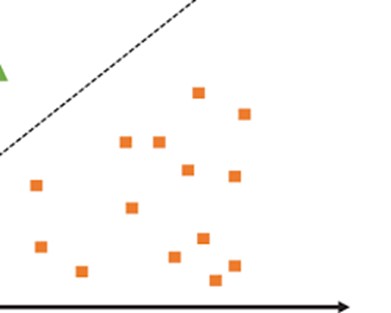
The study concluded that the ANFIS model, which combines the advantages of both ANN and FIS models, provided more accurate predictions of BTS compared to the individual ANN and FIS models. The R² values for the ANFIS, ANN, and FIS models were 0.92, 0.88, and 0.87, respectively. Additionally, the ANFIS model achieved the lowest RMSE value of 81.5%, compared to 87.5% for ANN and 89.5% for FIS. These results indicate that the ANFIS model is more effective in predicting BTS, offering improved accuracy and reliability over the other models evaluated in the study(Parsajoo et al., 2021).

In a study conducted by Miah M and Ahmed S., the authors investigated the use of machine learning to predict rock strength, specifically uniaxial compressive strength (UCS), from wellbore logging data. This approach holds significant value for the oil and gas industry, as it is more cost-effective than traditional methods and contributes to improved wellbore stability and optimized hydraulic fracturing. The input data for this study consisted of various wellbore log measurements, including acoustic travel time, which measures the time taken for a sound wave to travel through the rock, gamma ray, which measures the natural radioactivity of the rock, formation electron density, which assesses the density of electrons within the rock formation, and the shaliness effect, an indicator of the presence of clay minerals in the rock.



**Figure 9. Well logs**

Two machine learning models were employed in the study. The first was the Backpropagation Multilayer Perceptron (MLP) Artificial Neural Network (ANN), a complex algorithm inspired by the human brain and capable of learning intricate patterns from data. The second model was the Least Squares Support Vector Machine (LS-SVM) with Coupled Simulated Annealing (CSA) optimization. This model finds a hyperplane in a high-dimensional space to separate data points based on rock strength, with CSA optimization used to identify the most accurate hyperplane for prediction.

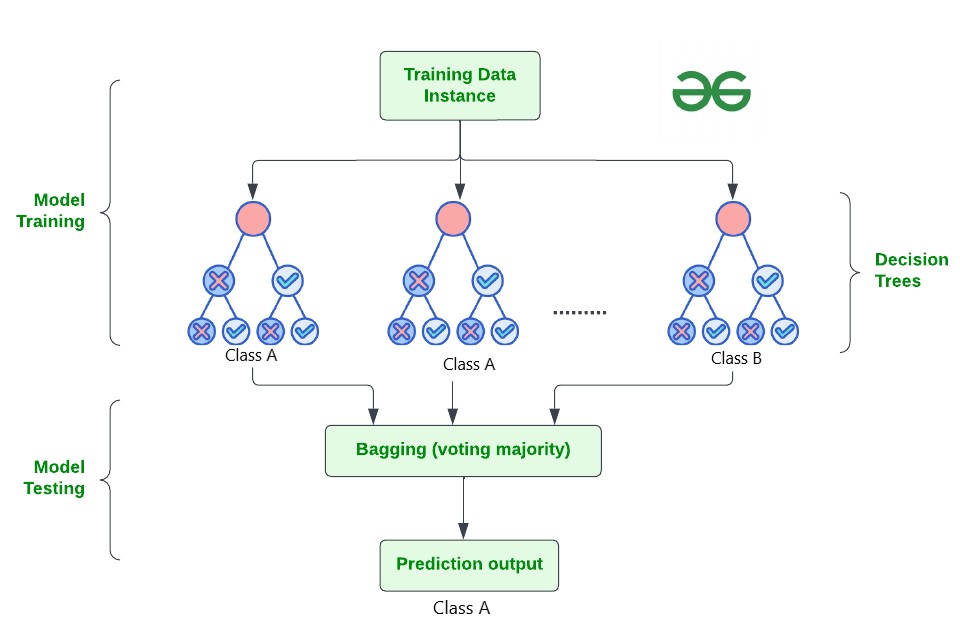


**Figure 10. Support Vector Machine**

The process involved several steps. First, the machine learning models were trained using the wellbore log data and corresponding measured rock strength values. Then, these trained models were used to predict rock strength for new data. The predicted rock strength was compared to measured values and existing log-based correlations for validation. The importance of each input parameter in the prediction process was also evaluated. The study concluded that both MLP-ANN and LS-SVM models achieved high accuracy in predicting rock strength, with predictions closely matching measured values. Acoustic travel time and gamma ray were identified as the most critical input parameters for rock strength estimation. Furthermore, new correlations for predicting rock strength were developed using the most significant log parameters. This machine learning approach enables the quick estimation of rock strength using wellbore logging data, reduces reliance on expensive laboratory measurements, and improves wellbore stability and the optimization of hydraulic fracturing in oil and gas exploration(Miah et al., 2020)*.*

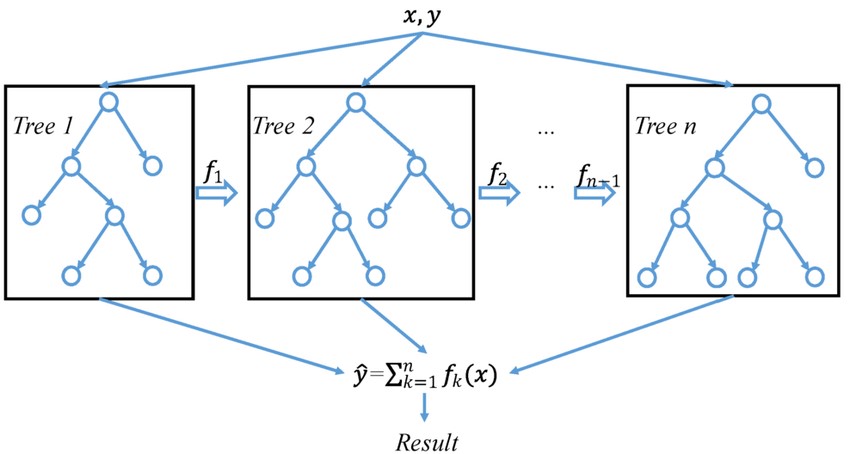
In a study conducted by Nasiri, Homafar, and Chelgani, the durability of travertine, a commonly used rock in construction and rock engineering, was assessed using machine learning algorithms to predict the uniaxial compressive strength (UCS) and modulus of elasticity (E).

The study highlighted the importance of UCS and Young's modulus of elasticity (E) in evaluating rock durability and classifying rocks based on their elastic deformation. These properties are essential for understanding stress and predicting failure points in rocks, which is essential for the safe construction of dams, mines, and fossil fuel projects. The input data used in this study included n-porosity, water absorption rate, dry unit weight, Schmidt Hammer rebound number, p-wave velocity, point load strength index, density, slake durability index, dry density, and Brazilian tensile strength. These variables were employed in the prediction models to estimate UCS and E. Several machine learning models were applied to this task. The Random Forest model, known for its ability to combine multiple decision trees to make accurate predictions, was utilized. This model improves accuracy and avoids overfitting by using a random subset of features and data points for each tree.



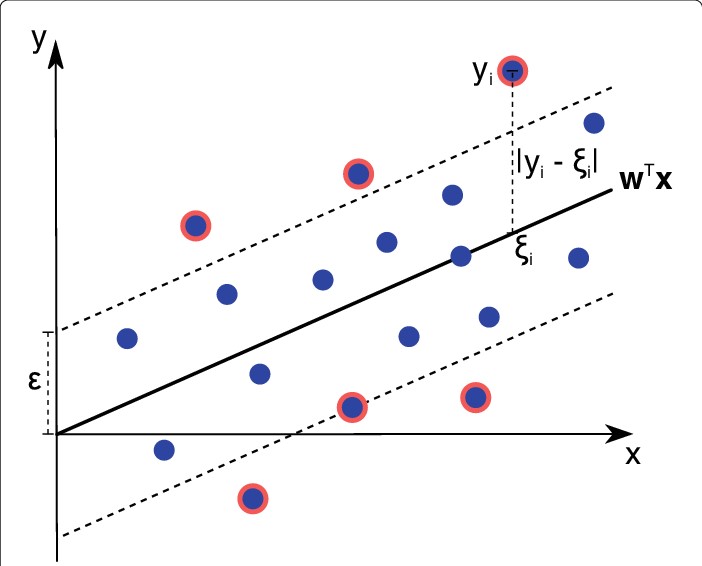
**Figure 11. Random Forest Algorithm**

Another model used was Extreme Gradient Boosting (XGBoost), a highly efficient machine learning algorithm known for its speed and accuracy. XGBoost focuses on optimizing each new model to improve overall accuracy, preventing overfitting and enhancing model generalization.



**Figure 12. General Architecture of XGBoost**

The third model employed was Support Vector Regression (SVR), which aims to find a hyperplane that best fits the training data while maximizing the margin around it, ensuring some tolerance for errors.



**Figure 13. Support Vector Regression**

Additionally, Shapley Additive Explanations (SHAP) were used as a technique to explain the predictions of the machine learning models, offering insights into how much each feature contributed to the model's output. The results and discussions focused on evaluating the effectiveness of various factors in predicting the mechanical properties of travertine rock, specifically UCS and E. The study used SHAP values to analyze the importance of different factors, revealing that p-wave velocity (Vp) had the strongest positive influence, while the point load index (PLI) had the weakest impact. The findings also confirmed that porosity negatively correlated with UCS and E, while a positive correlation existed between UCS/E and the point load strength index. When comparing the prediction accuracy of the machine learning models, XGBoost outperformed both Random Forest (RF) and Support Vector Regression (SVR). XGBoost's adaptability, reduced need for feature engineering, and efficient parallel processing were identified as significant advantages. The XGBoost model was used to generate predictive models for UCS and E, with 80% of the dataset used for training and the remaining samples for testing. Hyperparameters for XGBoost were optimized using a combination of default settings and trial and error through the Grid Search algorithm. The results demonstrated that UCS and E could be accurately predicted based on the rock properties. For validation purposes, the same dataset samples were used to generate RF and SVR models for UCS and E prediction. The XGBoost model was found to predict UCS and E with higher accuracy than the RF and SVR models. The study also noted that while both RF and XGBoost are ensemble methods, SVR is not. XGBoost, being a boosting technique, is more adaptable than RF, which is a bagging method. XGBoost allows for customization of the objective function, requires less feature engineering, and implements parallel processing, resulting in lower computational costs. In contrast, SVR performs well in high-dimensional spaces due to the kernel trick but has a higher computational cost. All three algorithms, however, were capable of handling outliers effectively. The conclusion of the study highlighted that Shapley Additive Explanations (SHAP) as an explainable artificial intelligence method could accurately assess correlations between variables and illustrate the effect of each rock property on the mechanical indexes (UCS and modulus of elasticity) of travertine. The SHAP values indicated that p-wave velocity (Vp) had the highest impact on E and UCS modeling, with Vp, Schmidt hammer rebound number, and point load index showing positive correlations with UCS and E, while porosity had a negative relationship. The evaluation of the models demonstrated that XGBoost provided extensive and accurate UCS and E prediction results, making it a powerful tool for modeling and managing problems in rock mechanics(Nasiri, Homafar and Chelgani, 2021)*.*

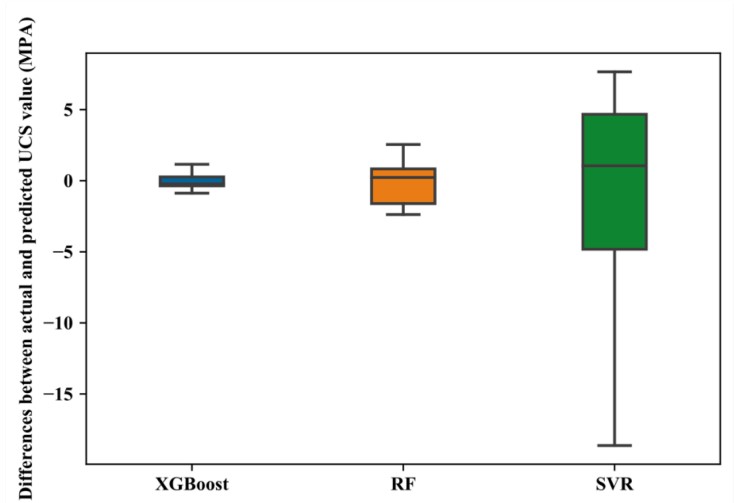


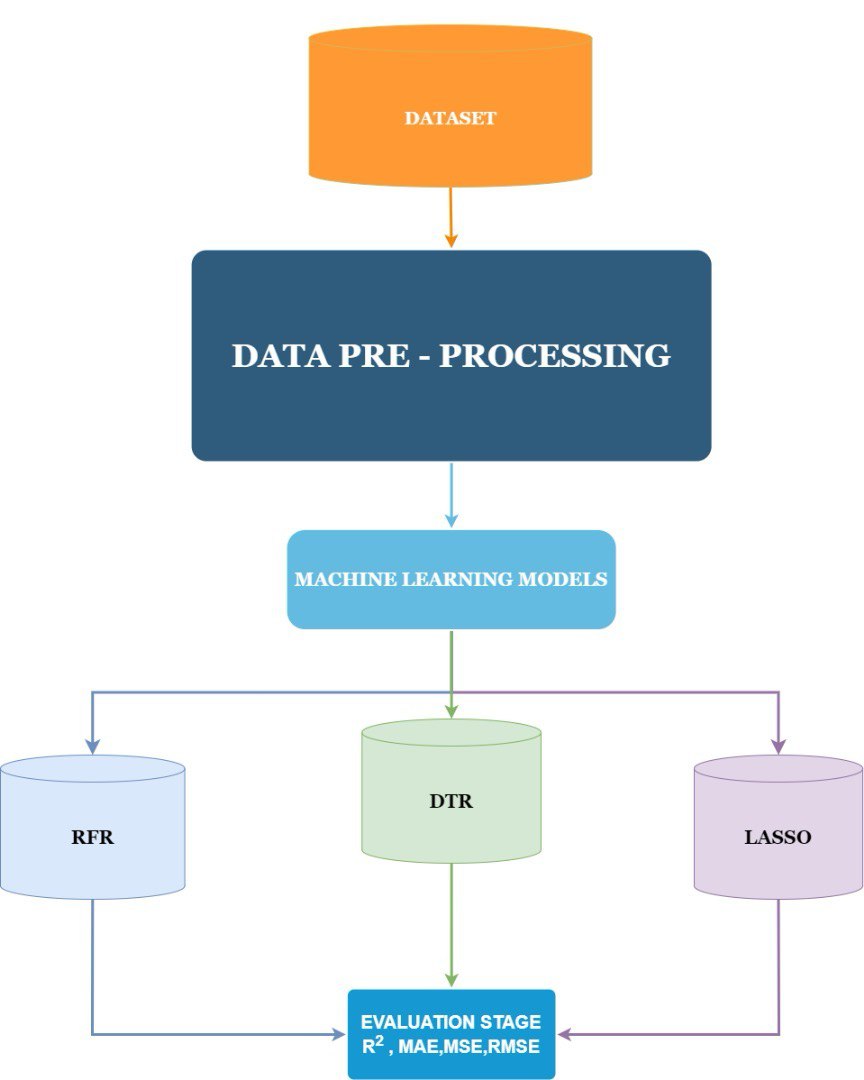
Figure 14. Box plot of actual and predicted UCS values

# CHAPTER 3

# METHODOLOGY

## 3.1 OVERVIEW

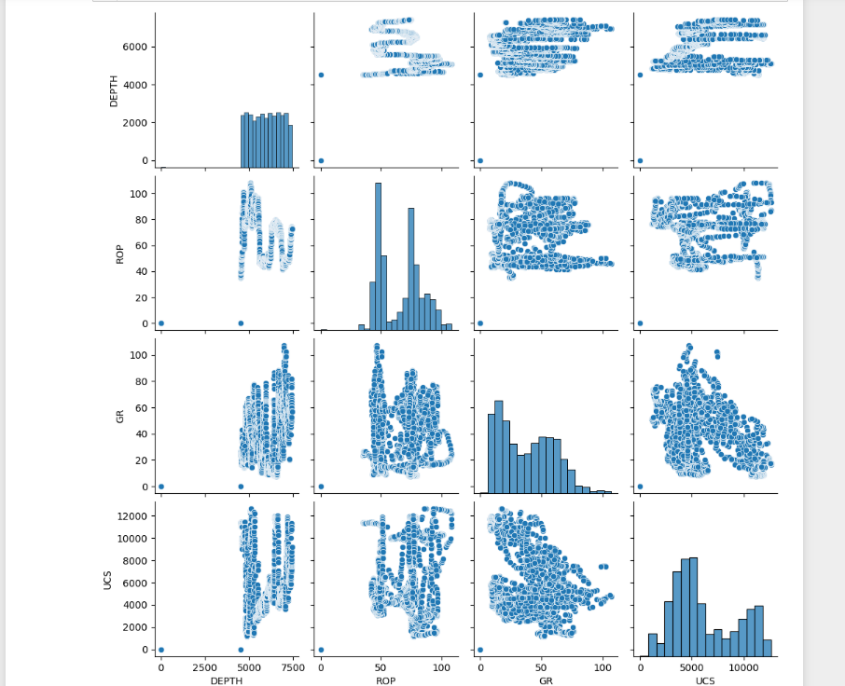
In this study, Python libraries were utilized to develop the mathematical procedures for LASSO Regression, Decision Tree and Random Forest. Several libraries used during the machine learning process include NumPy, Pandas, Scikit-learn (sklearn), Matplotlib or Seaborn. Collectively, these libraries created informative plots and aided in result interpretation. Decision Tree, Random Forest and Lasso Regression were widely utilized in predictions due to their adaptability, capability to model non-linear relationships, and ensemble learning technique. These algorithms provided valuable insights into feature importance and could efficiently handle complex datasets. Their capacity to deliver precise predictions and interpret results enhances decision making regarding production optimization and resource allocation. Fig. 1 illustrates the architecture of the fundamental system construction procedures. The suggested framework consists of a series of stages, which include Data Set Description, Data Preprocessing, Implementation of Machine Learning models and Evaluation stage.



**Figure 15 - General Workflow**

## 3.2 DATA PREPROCESSING

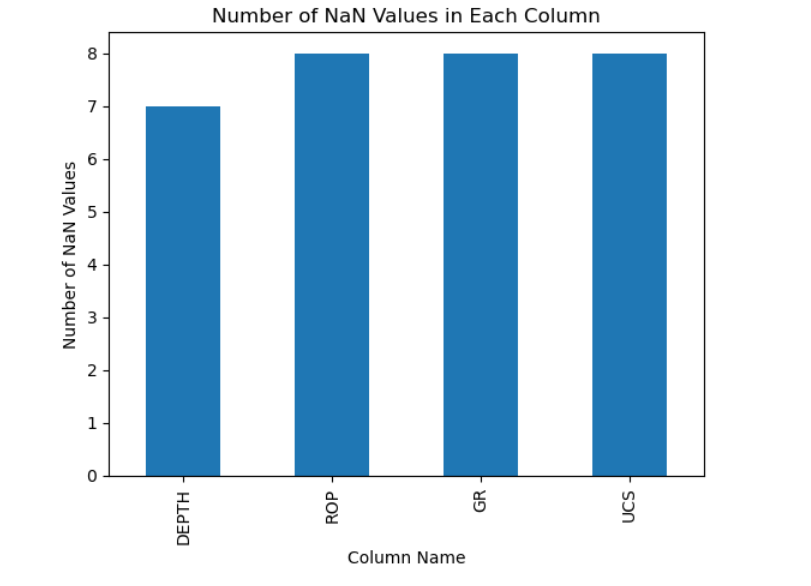
The first and most critical stage of the research is collection of data. This research used Jupyter Notebook as an environment for programming in Python. The dataset included Depth, Gamma Ray, Rate of Penetration and UCS. It was represented in 2066 rows and 4 columns. Figure 2 provides part of the data set, specifically showcasing the data trends for DEPTH, GAMMA Ray, Rate of Penetration and UCS.



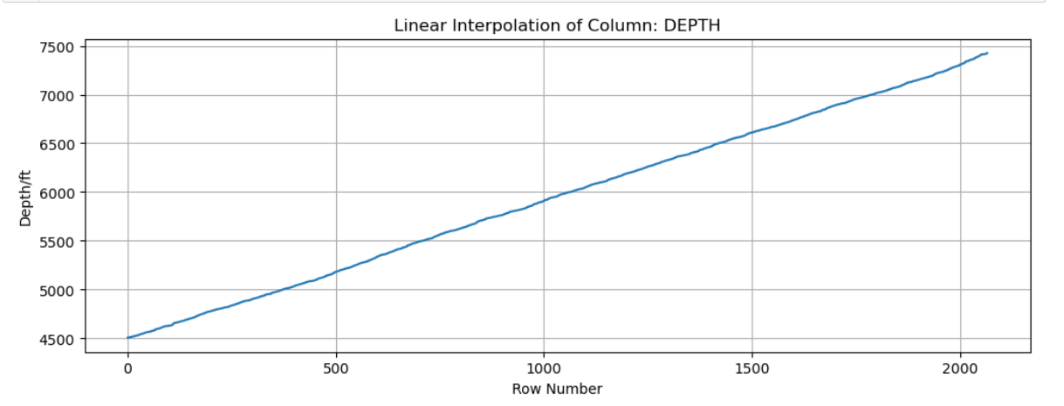
**Figure 16 - Trend in Data**

### 3.2.1 Data Cleaning

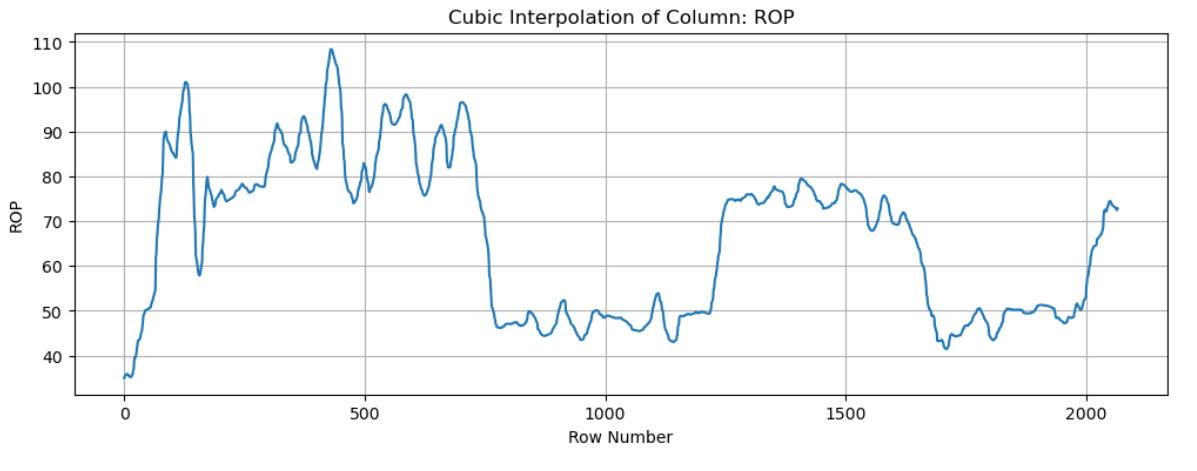
An in-depth investigation was conducted on the dataset to gain insights into its structure, patterns, and relationships among variables. By using the find and replace approach (CTRL+H) , the missing values were represented as “NaN” (Not a Number) to ensure uniformity in the representation of NaN values due to the fact that some were represented as “Nan”. The Depth Column had 7 NaN values, which represents approximately 0.34% of the dataset. The other parameters, GR, ROP, and UCS, each had 8 NaN values, representing approximately 0.39% of the dataset. NaN values are often due to sensor failure during data recording process. Data was cleaned and prepared for analysis by handling missing values and outliers which were represented as NaN. The method used for data cleaning was Interpolation. Interpolation is a useful technique in data cleaning because it allows for the estimation of missing data points by using the values of other data points in a dataset. This helped to create a more complete and continuous dataset, which was useful for visualization, analysis, and modeling. One advantage of interpolation is that it can help to preserve the overall trend and structure of the data. By estimating missing values based on nearby data points, interpolation can help to maintain the underlying pattern of the data. Interpolation was performed for each column ie, DEPTH, Gamma Ray, Rate of Penetration (ROP) and Uniaxial Compressive Strength ( UCS). Linear interpolation was used for the DEPTH column as the relationship between the DEPTH values and the Row Numbers were linear. Cubic interpolation was used for GR, ROP and UCS. This is because there was a monotonic relationship (a consistent pattern of decrease and increase) between the row numbers and their values.



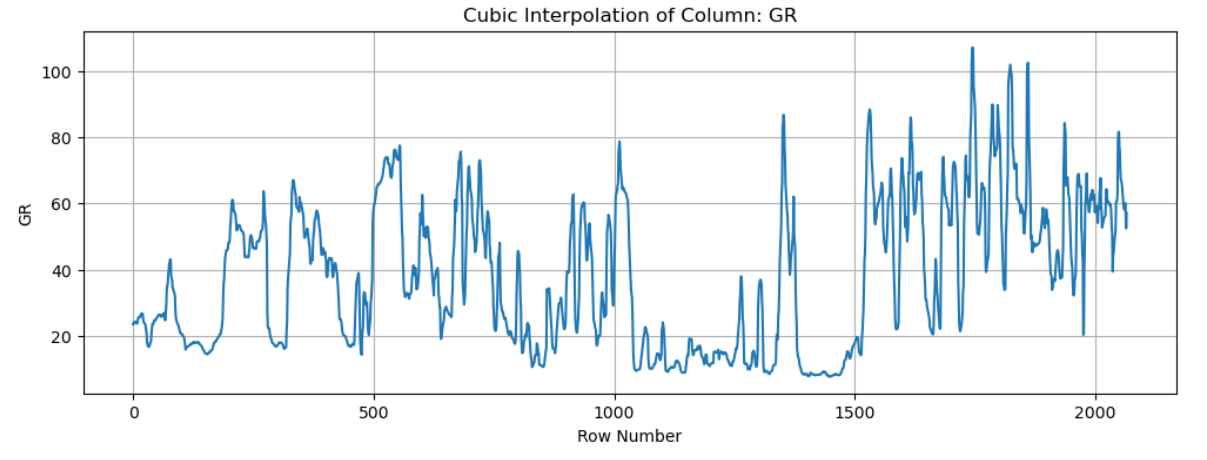
**Figure 17 - Null Values in DataFrame**



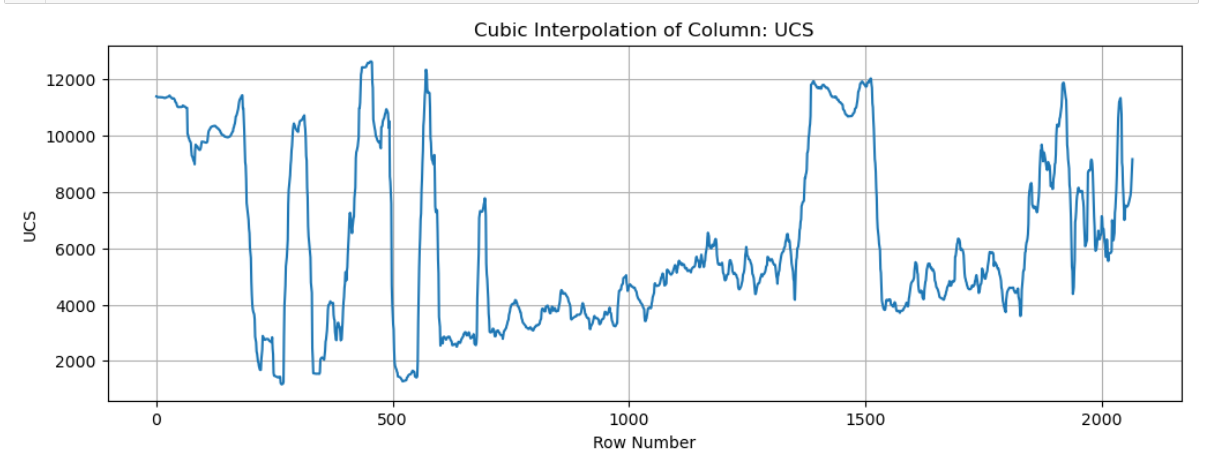
**Figure 18 - Linear Interpolation of Depth Column**



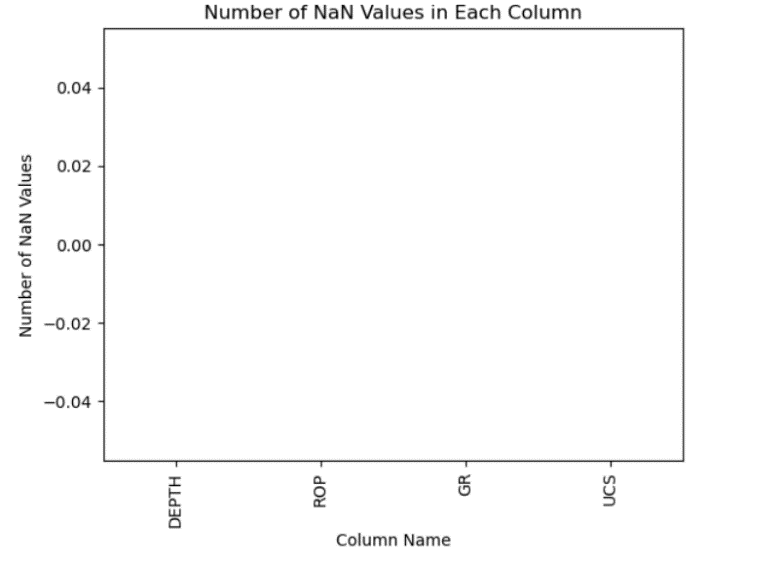
**Figure 19 - Interpolation of ROP column**



**Figure 20 - Interpolation of GR column**



**Figure 21 - Interpolation of UCS Column**



**Figure 22 - Visualizing Clean Data**

### 3.2.2 Feature Scaling

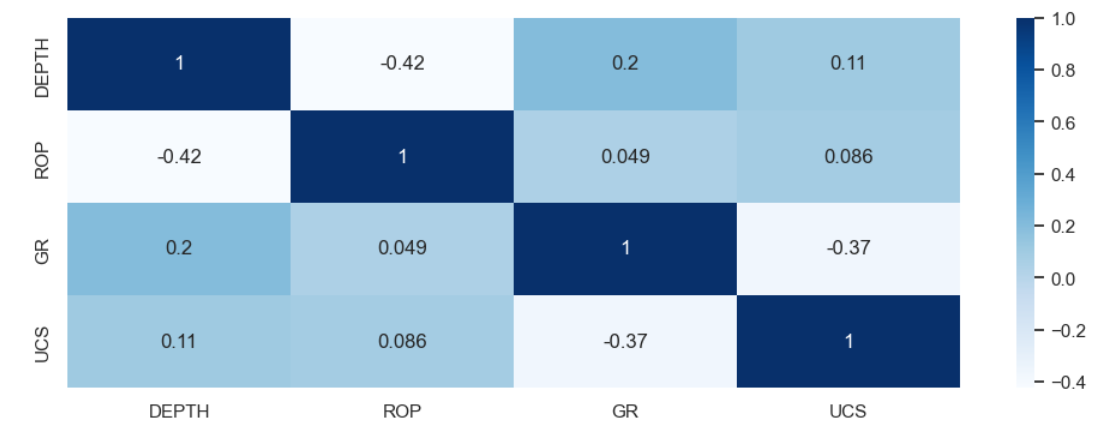
The method of standardization was used to scale the data. The purpose of standardization, also known as feature scaling, is to transform the features (independent variables) of a dataset so that they have a mean of 0 and a standard deviation of 1. This process ensures that all features are on the same scale, which would be beneficial for the machine learning algorithms. Standardization ensures that all features contribute equally to the model's decision-making process. Without standardization, features with larger scales may dominate the training process. Lasso Regression which employs regularization techniques such as L1 regularization assumes that all features would be on the same scale to prevent overfitting.

For each feature, the mean and the standard deviation which is a measure of the spread is calculated across all the samples in the dataset. The mean is subtracted from each feature value to centre the data around zero. Each centred value is divided by its standard deviation. This scales the data to have a standard deviation of 1. The ‘StandardScaler’ class from the ‘sklearn.preprocessing’ module is used for standardization. Mathematically, standardization which is conventionally known as Z-score is represented by:

Where X represents a specific value in a feature, represents the mean and represents the standard deviation of the particular feature.

### 3.2.3 Correlation Matrix and Heat Map

Correlation matrix is a square matrix that contains correlation coefficients between pairs of variables in a dataset. It measures the strength and direction of the relationship between two variables, whether linear or monotonic. The method used to estimate the correlation coefficients was Spearman’s method. This was chosen because Spearman’s correlation coefficient is suitable when there is a monotonic relationship between the variables. The pair plots from Seaborn indicated a monotonic relationship, which solidified the decision to use Spearman’s method to compute the correlation coefficients. Figure 8 shows the heatmap of correlation coefficients for visualization purposes.



**Figure 23 - Correlation Matrix and Heatmap**

The correlation coefficient of 0.11 between Depth and UCS of ROCK suggests a very weak positive relationship, indicating that as Depth increases, UCS tends to increase slightly. However, this weak correlation implies that Depth is unlikely to have significant predictive power for UCS, but it may still contribute a small amount of weight to the model during training.

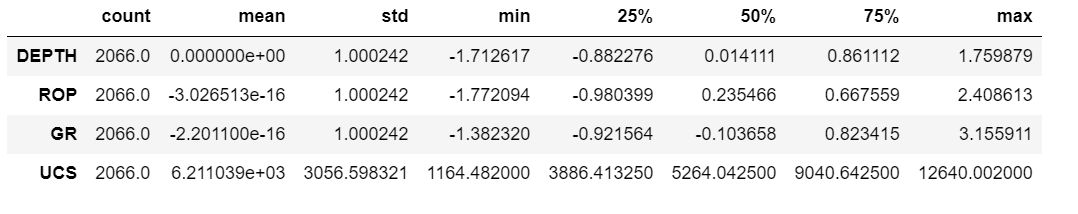
Similarly, the correlation coefficient of 0.086 between Depth and ROP indicates a very weak positive relationship, suggesting that as Depth increases, ROP tends to increase slightly. Despite this minimal correlation, Depth is unlikely to have meaningful predictive power for ROP, but it may still have a small impact on the model during training.

Last but not least the Gamma Ray Index has a correlation coefficient of -0.37 with UCS, indicating a moderate negative relationship. This means that as the Gamma Ray Index increases, UCS tends to decrease. In the context of petroleum formation evaluation, a higher Gamma Ray Index often signifies higher shale content, which is associated with weaker rock formations. Therefore, rocks with higher shale content (higher Gamma Ray Index) tend to have lower UCS, explaining the negative relationship.

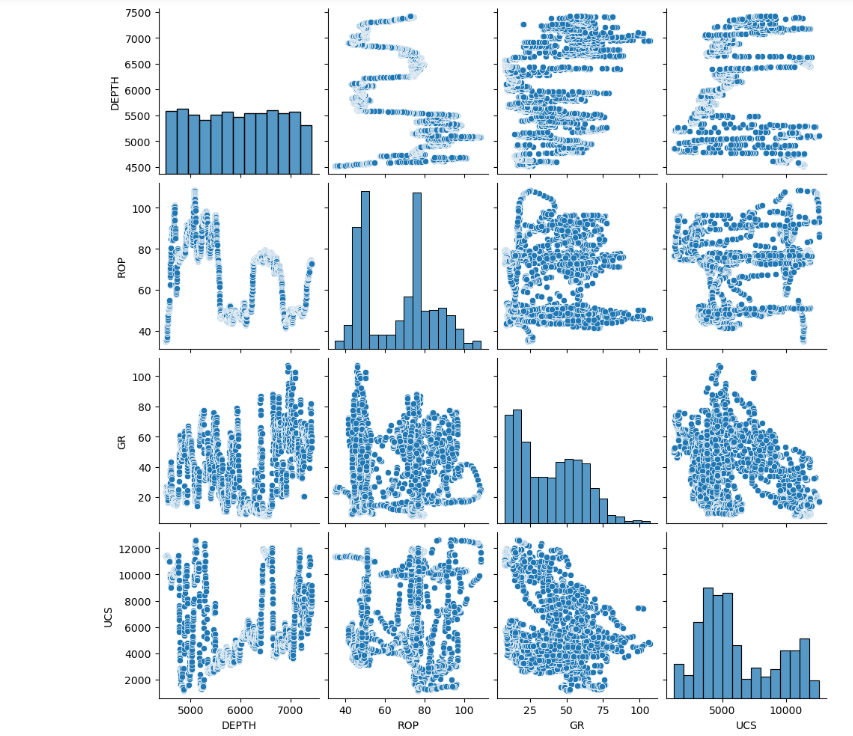
## 3.3 EXPLORATORY DATA ANALYSIS

### 3.3.1 Statistical Measures

The median UCS value is 5264.042. This indicates that 50% of the UCS values are below this point, while the other 50% are above. The median provides a central value around which the UCS data is distributed. The UCS values range from a minimum of 1164.482 to a maximum of 12640.002. This wide range highlights the significant variability in the UCS values, indicating that the data includes both very low and very high values. The 75th percentile value is 9040.643. This means that 75% of the UCS values are below 9040.643, while the top 25% are above this value. The high 75th percentile suggests that a significant portion of the data points have very high UCS values.



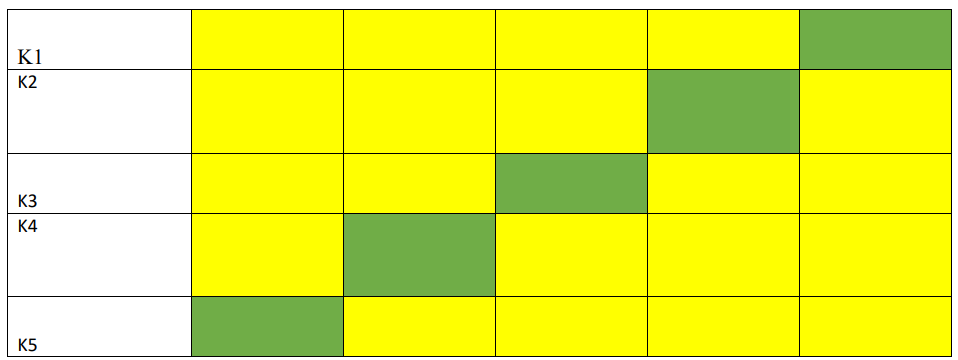
### 3.3.2 Data Visualization



### 3.3.3 Data Splitting

This is the last stage in exploratory data analysis. Cross validation technique was used to evaluate the performance of a model on unseen data. It involvesdividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Cross validation is an important step in the machine learning process and helps to ensure that the model selected for deployment is robust and generalizes well to new data. The shape of the training and test data frames varied with each iteration of the K-fold cross validation.

The first iteration had 1652 samples in the training set with 3 features each, and 414 samples in the test set with 3 features each after employing cross validation of 5 folds. The data was split into 5 folds to allow for a fair number of datasets to be introduced to the test data. The target variable (UCS) for the training set had 1652 elements, and the target variable for the test set has 414 elements



**Figure 24 - Data Splitting Architecture**

## 3.4 EVALUATION METRICS

Evaluation metrics are quantitative measures used to assess the performance of a model . The performance of a regression model is assessed using data, often known as testing data, using the following measures.

### 3.4.1 Mean Absolute Error ( MAE)

The Mean Absolute Error (MAE) is useful when dealing with data that may contain outliers affecting the values. MAE is less sensitive to outliers during training, providing a balanced and inclusive performance measure for the model. On the other hand, if the test set includes numerous outliers, the model's performance will be enhanced. This is illustrated in Equation (1) where the ideal MAE value is 0, and the worst possible value is positive infinity.

|  |  |
| --- | --- |
|  | (3) |

### 3.4.2 Mean Square Error

The use of MSE (Mean Squared Error) is particularly suitable for identifying outliers. This is because MSE effectively assigns greater significance to extreme data points. When the model makes a poor prediction, the squaring component of the MSE function significantly magnifies the error's impact as demonstrated in Equation 2.

|  |  |
| --- | --- |
|  | (4) |

### 3.4.3 Coefficient of Determination

Linear regression models employ a well-defined statistic called the coefficient of determination (R²) to quantify the extent to which the observed variation in the dependent variable can be attributed to known predictors [41]. R², also known as the goodness of fit, ranges from 0 to 1. An R² value of 1 indicates that the model explains all the variability of the response data around its mean, while an R² of 0 indicates that the model explains none of the variability. It is calculated using the following Equation 3.

|  |  |
| --- | --- |
|  | (5) |

# CHAPTER 4

# RESULTS AND DISCUSSION

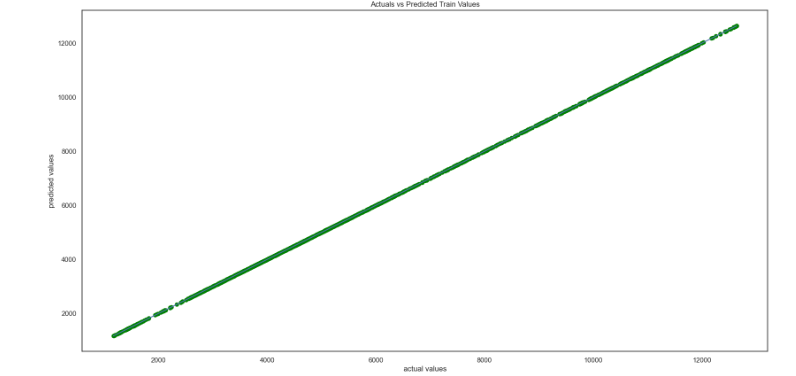
## 4.1 Decision Tree Model

Decision Tree algorithms, known for their interpretability and effectiveness in classification and regression tasks, stand as foundational tools in the realm of supervised learning. A decision tree is a type of machine learning model that is used for both classification and regression tasks. It's called "non-parametric" because it makes very few assumptions about the form of the underlying data distribution. Instead of assuming a specific functional form (like linear regression does), decision trees adaptively partition the feature space into smaller regions, making them flexible and capable of capturing complex relationships between features and target variables. The structure of a decision tree entails root nodes/decision node, branches, internal nodes and leaf nodes. The hierarchical nature of decision trees means that decisions are made in a top-down manner, starting from the root node and progressing down to the leaf nodes. Each internal node represents a decision based on a feature, leading to a split in the dataset, and each leaf node represents a final decision or prediction. For each internal node, the algorithm repeats the process of selecting the best feature to split on. This step is repeated recursively for each subset, creating a tree structure where each internal node represents a decision based on a feature, and each branch represents the outcome of that decision.

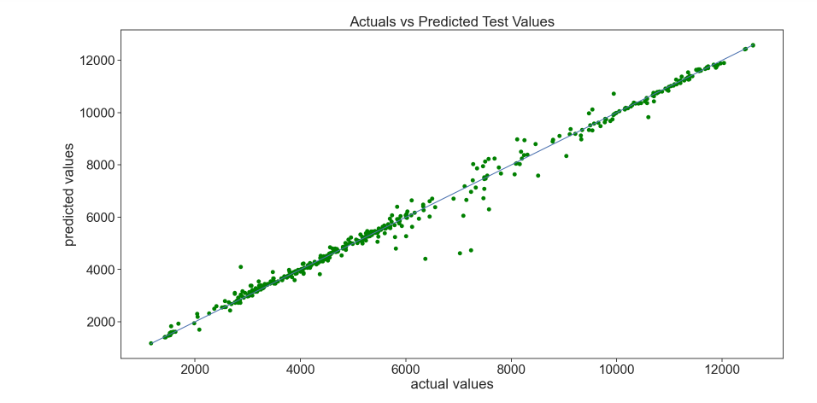
The Decision Tree model performs perfectly on the training data with an R² of 1.0 and zero errors, indicating the model perfectly fits the training data. One disadvantage of the Decision Tree model is that it is sensitive to small variations in data, leading to different tree structures and potential instability.

Table 1 - Evaluation Metrics for Decision Tree Model

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE MODEL | | | |
| TRAIN DATA | | **TEST DATA** | |
| MSE | 0 | **MSE** | 105,025.82 |
| RMSE | 0 | **RMSE** | 324.08 |
| MAE | 0 | **MAE** | 138.53 |
| R2 | 1 | **R2** | 0.99 |



**Figure 26 - Train Data Plot for Decision Tree Model**



**Figure 27 – Test Data Plot for Decision Tree Model**

## 4.2 Random Forest Model

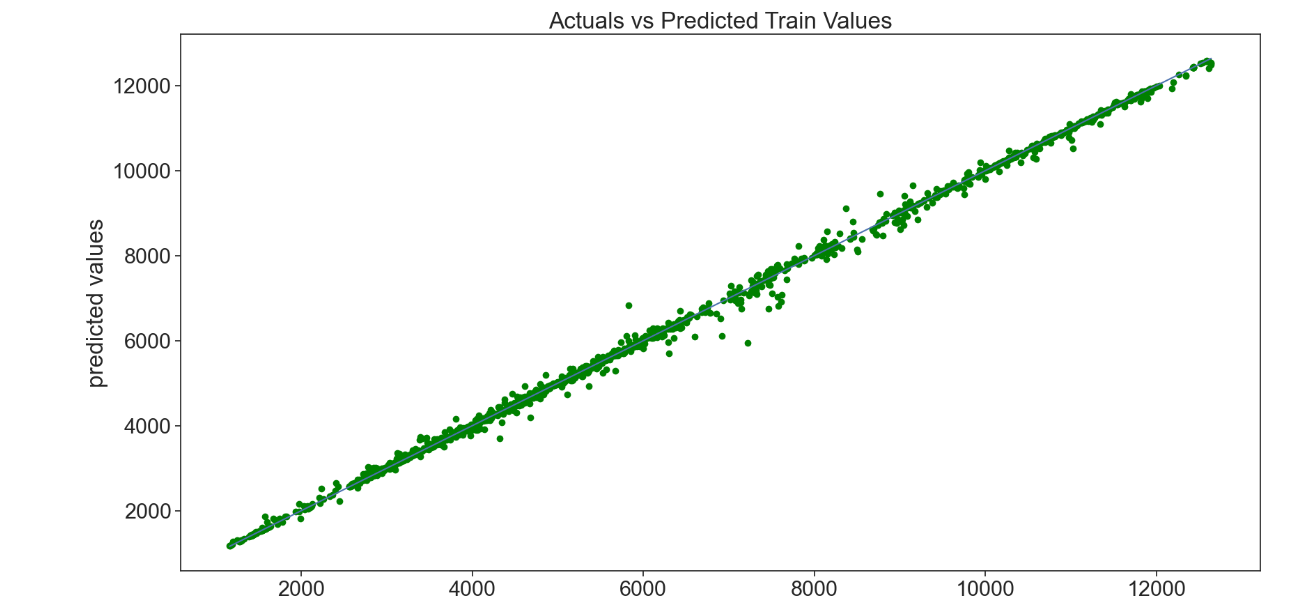
The Random Forest algorithm is a widely used ensemble learning method that combines the predictions of multiple decision trees to enhance the overall accuracy and robustness of the model. It is particularly well-suited for handling complex datasets and capturing non-linear relationships between variables. By aggregating the results of numerous decision trees, Random Forest reduces the risk of overfitting and provides a more generalized model that performs well on unseen data.

In this project, a Random Forest model was implemented to predict outcomes based on a given dataset. The model’s performance was evaluated using various statistical metrics to understand its effectiveness and reliability. The following sections will provide a comprehensive discussion of the model's performance, highlighting its strengths and the insights gained from the results.

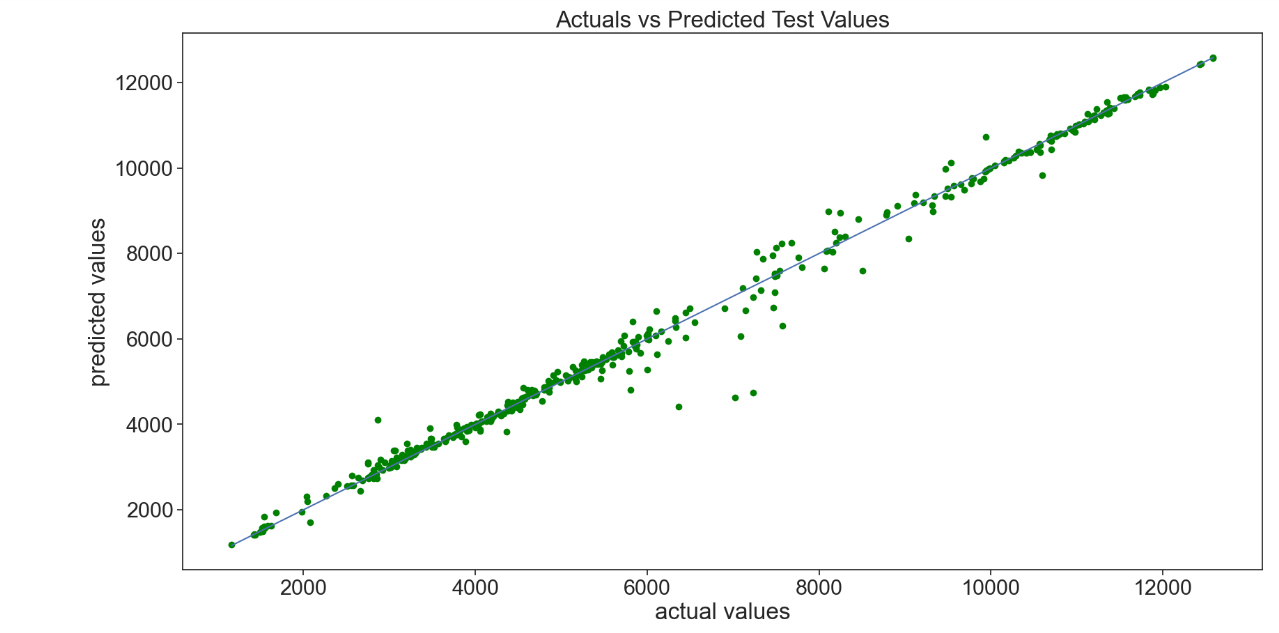
Mean Square Error of 11,068.49 on the training set suggests that the model fits the training data very closely, with relatively small average squared errors. However, MSE of 133,000.32 on the test set indicates a significant increase in error, suggesting that the model's performance on new, unseen data is much worse. This large gap between training and test MSE is a strong indicator of overfitting. An R² of 0.998805 on the training set implies that the model explains approximately 99.88% of the variance in the training data. This near-perfect score suggests that the model has almost completely captured the patterns in the training set. An R² of 0.986212 on the test set, while still high, shows a slight decrease compared to the training R². This drop, combined with the increased error metrics (MSE, RMSE, MAE), indicates that while the model explains a large portion of the variance in the test data, its performance is not as robust on unseen data.

Table 2 - Evaluation Metrics for Random Forest Model

|  |  |  |  |
| --- | --- | --- | --- |
| RANDOM FOREST MODEL | | | |
| TRAIN DATA | | **TEST DATA** | |
| MSE | 11068.49 | **MSE** | 133000.32 |
| RMSE | 105.21 | **RMSE** | 364.69 |
| MAE | 51.28 | **MAE** | 155.25 |
| R2 | 0.998 | **R2** | 0.986 |



**Figure 29 - Train Data Plot for Random Forest**

****

**Figure 30 - Test Data Plot for Random Forest**

## 4.3 Lasso Regression

The Lasso Regression model was evaluated using four key metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R2). These metrics help to assess the model's performance on both the training and test datasets. The MSE values for the training and test datasets are 7,383,632.68 and 7,579,549.16, respectively. These high MSE values indicate that the model's predictions are significantly off from the actual values, which suggests that the model is not performing well. The similarity in MSE values between the training and test sets shows that the model has consistent performance across both datasets but is still inaccurate.

The RMSE values are 2,717.284 for training and 2,753.098 for testing. RMSE is the square root of MSE, and it also indicates the model's prediction error. The values show that on average, the model's predictions are off by around 2,700 units, further indicating the model's lack of precision.

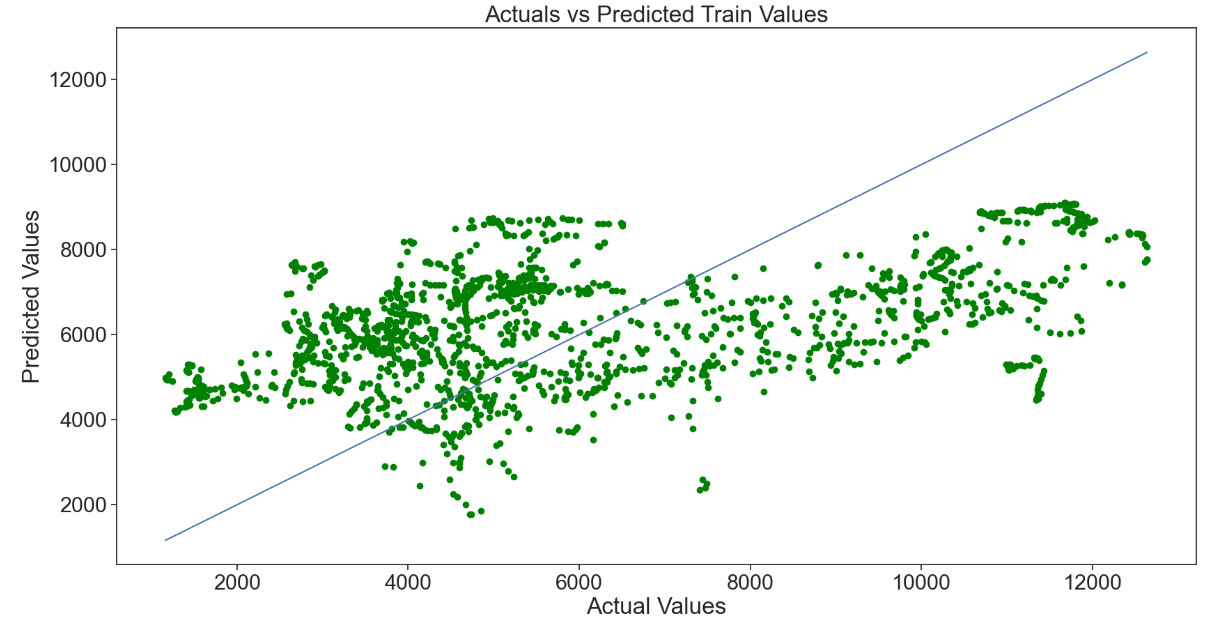
The MAE values are 2,375.926 for the training set and 2,417.697 for the test set.

MAE, which measures the average absolute difference between predicted and actual values, aligns with the RMSE findings. It suggests that, on average, the model's predictions deviate by around 2,400 units from the actual values.

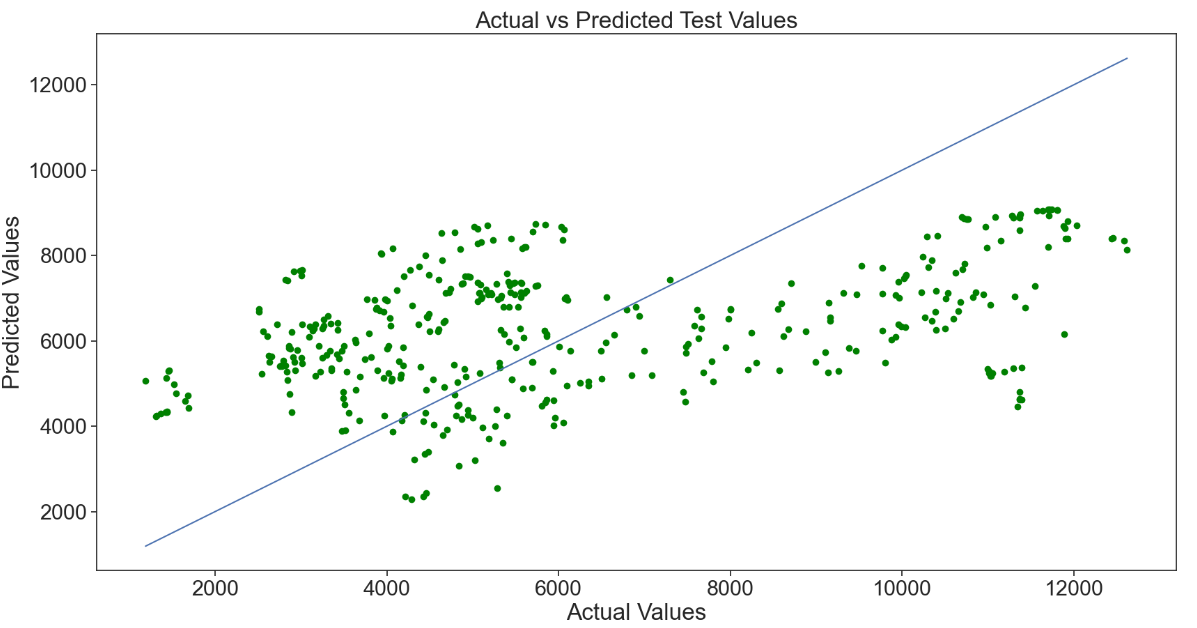
The 𝑅2 values are 0.203 for the training set and 0.214 for the test set. R2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. The values close to zero suggest that the model explains only about 20-21% of the variance in the data, indicating significant underfitting.

Table 3 – Evaluation metrics for LASSO regression model

|  |  |  |  |
| --- | --- | --- | --- |
| LASSO REGRESSION | | | |
| TRAIN DATA | | **TEST DATA** | |
| MSE | 7,383,632.68 | **MSE** | 7,579,549.2 |
| RMSE | 2717.28 | **RMSE** | 2753 |
| MAE | 2375.926 | **MAE** | 2417.6 |
| R2 | 0.203 | **R2** | 0.214 |



**Figure 32 - Train Data Plot for Lasso Regression**

****

**Figure 33 - Test Data Plot for Lasso Regression**

# CHAPTER 5

# CONCLUSION AND RECOMMENDATIONS

# 5.1 CONCLUSION

In this project, we investigated the prediction of rock strength using Lasso Regression, Decision Tree, and Random Forest machine learning algorithms. The coefficient of determination (R²) for all the models was in the range of 97-99%, indicating a high correlation between the predicted and actual values. However, the Mean Squared Error (MSE) values were very large, signifying that the models are weak and not robust. The high R² values can be misleading, as they do not necessarily indicate a good model performance. The large MSE values suggest that the models are not able to accurately predict the rock strength. All three models (Lasso Regression, Decision Tree, and Random Forest) performed poorly, indicating that the relationships between the input features and rock strength are complex and cannot be captured by these models. Therefore, it can be concluded that the models are weak and not suitable for predicting rock strength. Further research is needed to develop more robust and accurate models.

# 5.2 RECOMMENDATION

Following a comprehensive analysis of the project's machine learning models, it is recommended that hyperparameter tuning be conducted to enhance their predictive performance. Hyperparameter tuning is a vital step in machine learning, as it enables the optimization of model parameters to achieve the best possible performance on a given dataset. By fine-tuning hyperparameters, models can adapt to the underlying patterns and relationships in the data, leading to improved accuracy, robustness, and generalizability. In the context of this project, hyperparameter tuning presents a key opportunity to enhance the predictive power of the Random Forest, Lasso Regression, and Decision Tree models. Optimizing hyperparameters could allow these models to better capture the nuances of the data, reduce errors, and improve overall performance.

1. For the Random Forest Model, a systematic grid search approach is recommended to tune key parameters such as the number of estimators, maximum depth of the trees, and minimum samples per leaf. It is anticipated that this optimization process will improve the model's ability to generalize to unseen data while achieving a balanced trade-off between bias and variance.

2. In the case of the Lasso Regression Model, it is suggested that the regularization parameter (alpha) be optimized using cross-validation techniques. The selection of the optimal alpha value is expected to minimize prediction errors while promoting sparsity in the model coefficients. This process may lead to a more interpretable and effective model while enhancing predictive accuracy and reducing overfitting through the penalization of less informative features.

3. Lastly, for the Decision Tree Model, it is advised that hyperparameters including maximum depth, minimum samples per leaf, and splitting criteria be explored. The use of grid search and randomized search strategies is encouraged to identify the best combination of parameters.

This hyperparameter tuning is expected to tailor the models to capture complex relationships within the data while preventing excessive branching and overfitting.

# REFERENCES

Dusim, H.H. (2020) ‘Statement of Problem Template’, *I-JaMCSIIX 2020*, (September), pp. 3–5. Available at: https://www.researchgate.net/publication/354309875.

Liu, Q. *et al.* (2015) ‘Cold Regions Science and Technology A prediction model for uniaxial compressive strength of deteriorated rocks due to freeze – thaw’, *Cold Regions Science and Technology*, 120, pp. 96–107. Available at: https://doi.org/10.1016/j.coldregions.2015.09.013.

Manoharan, A. (2024) ‘Machine Learning Algorithms for Personalized Learning Paths’, International Research Journal of MoManoharan, A. (2024). Machine Learning Algorithms for

Personalized Learning Paths. International Research Journal of Modernization in Engineering

Technology and Science, March. https://doi.org/10.56726/irjmets49965derni [Preprint], (March). Available at: https://doi.org/10.56726/irjmets49965.

Naeimipour, A. *et al.* (2018) ‘International Journal of Rock Mechanics and Mining Sciences Estimation of rock strength using scratch test by a miniature disc cutter on rock cores or inside boreholes’, *International Journal of Rock Mechanics and Mining Sciences*, 107(May), pp. 9–18.

Available at: https://doi.org/10.1016/j.ijrmms.2018.03.020.

Nasiri, H., Homafar, A. and Chelgani, S.C. (2021) ‘Results in Geophysical Sciences Prediction of uniaxial compressive strength and modulus of elasticity for Travertine samples using an explainable artificial intelligence’, *Results in Geophysical Sciences*, 8(October), p. 100034. Available at: https://doi.org/10.1016/j.ringps.2021.100034.

Parsajoo, M. *et al.* (2021) ‘Transportation Geotechnics Tensile strength prediction of rock material using non-destructive tests : A comparative intelligent study’, *Transportation Geotechnics*,

31(May), p. 100652. Available at: https://doi.org/10.1016/j.trgeo.2021.100652.

Weng, M.C. *et al.* (2024) ‘Influence of grain contents on the dynamic strength of rock-like

materials’, *Engineering Geology*, 331(November 2023), p. 107456. Available at:

https://doi.org/10.1016/j.enggeo.2024.107456.

Xiong, X. *et al.* (2021) ‘Intelligent Prediction Model of the Triaxial Compressive Strength of Rock

Subjected to Freeze-Thaw Cycles Based on a Genetic Algorithm and Artificial Neural Network’, 2021.

Yuan, C. *et al.* (2024) ‘Research on strength prediction of crack rock mass based on random forest

algorithm’, *Bulletin of Engineering Geology and the Environment*, 83(4). Available at: https://doi.org/10.1007/s10064-024-03629-6.

*et al.* (2024) ‘Classifying the Supervised Machine Learning and Comparing the Performances of the Algorithms’, *International Journal of Advanced Research*, 12(01), pp. 422–438. Available at: https://doi.org/10.21474/ijar01/18138.