Application of Machine learning in predicting the yield strength of API steels

Gaurav Kumar*, Qasim Murtaza, Ashutosh Bagchi, G. Kumar, Darshan Lal

Department of Mechanical Engineering, Delhi Technological University, New Delhi 110042

Email*: gauravkumar2k19.me090@gmail.com

Abstract:

Machine learning has become increasingly important in various fields, including the mechanical industry. API steels being a subgroup of high-strength low-alloy (HSLA) steels have been designed for use in the petroleum industry. In this research, the application of machine learning models to estimate the mechanical properties of API steels was explored. Both non-linear and linear machine learning models were employed to predict the yield strength of API steels. The models were evaluated using different performance metrics on test samples, which produced promising results. Random forest model proved to be effective in estimating the yield strength of API steels with a R2 Score of 0.95. The results exhibit the effectiveness of machine learning techniques in predicting mechanical properties, making them a valuable tool for researchers and engineers in the materials industry.

Keywords: Machine Learning, API steels, Data analysis, Yield Strength

1. Introduction:

API steels are a type of high-strength low-alloy (HSLA) steel [1] primarily developed for use in the petroleum industry [2]. A mixture of alloying elements, such as chromium, molybdenum, and nickel are present in these steels, which impart enhanced strength and corrosion resistance to them [3]. Their design is specifically aimed at enduring extreme temperatures, pressures, and corrosive environments [4,5] that are frequently encountered in the oil and gas sector. API steels are graded accordingly to maintain the quality of the steels used for the production of pipes and other equipment. The most widely used API grade classifications are API 5L for seamless and welded pipes, API 5CT and API 5B for casing and tubing, and API 5D for drill pipe [6]. Given their exceptional strength and toughness, API steels are extensively used in various applications, such as pipelines, offshore platforms, and drilling equipment, to withstand challenging industry conditions.

Machine learning, a branch of artificial intelligence, enables computers to learn from data and make decisions or predictions without human intervention. The three primary types of machine learning include reinforcement learning, unsupervised learning, and supervised learning. In supervised learning, labeled data is used to teach the algorithm in order to predict new data. Unsupervised learning entails discovering relationships and patterns in unlabeled data, while reinforcement learning involves learning through trial and error in a dynamic environment. Machine learning has widespread applications, including fraud detection, speech and image recognition, autonomous vehicles, predictive maintenance, natural language processing, and recommendation systems. With the continuous growth in data generation, machine learning is gaining importance in making sense of and extracting insights from the vast amount of data [7].

The utility of machine learning in production engineering and material science is diverse and extensive. One of its primary applications is the optimization of manufacturing processes, which is achieved through the prediction of defects, monitoring of equipment performance, and enhancement of product quality. In material science, machine learning techniques are used to expedite the discovery of new materials, predict their properties and performance, and optimize their processing [8]. Furthermore, machine learning algorithms are leveraged for predicting the fatigue life of materials [9], analyzing fracture patterns [10], and detecting material defects [11]. Additionally, predictive maintenance is another area where machine learning is used to identify potential equipment failures before they occur, thereby ensuring uninterrupted operations and reducing downtime [12].

This research aims to devise and compare several machine learning models to predict the minimum yield strength (in ksi) of API steels based on their mechanical properties and chemical composition. The dataset consisted of API steels with confirmed target values of yield strength, which were used to train and evaluate the models. The objective of this study is to determine the most precise and dependable ML model to anticipate the yield strength of API steels.

2. Methodology:

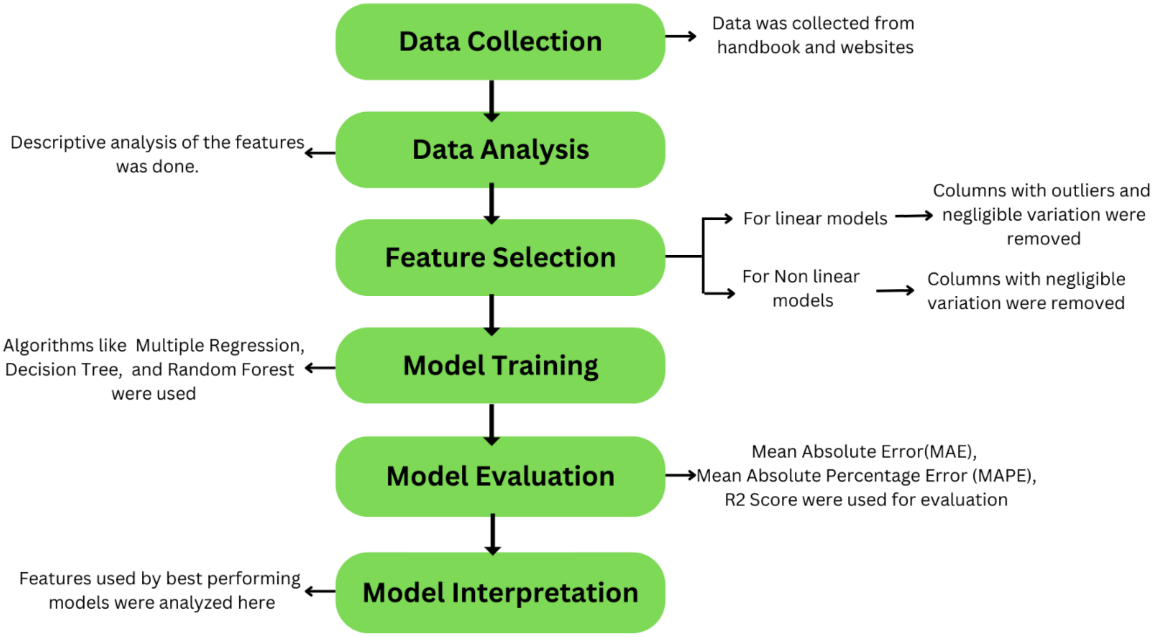
Data Collection
Data was collected from handbook and websites
Data Analysis
Descriptive analysis of the features was done.
Feature Selection
For linear models
Columns with outliers and negligible variation were removed
For Non linear models
Columns with negligible variation were removed
Model Training
Algorithms like Multiple Regression, Decision Tree, and Random Forest were used
Model Evaluation
Mean Absolute Error(MAE), Mean Absolute Percentage Error (MAPE), R2 Score were used for evaluation
Model Interpretation
Features used by best performing models were analyzed here

Figure 1: Flowchart of the steps followed in the analysis

This study utilized machine learning techniques to estimate the yield strength of API steels. As depicted in Figure 1, Initially a dataset of API steels was gathered, and their chemical composition, physical properties, and mechanical properties were recorded. Since the variation in physical properties was negligible, they were eliminated from the analysis. Subsequently, the dataset was divided into two sets, one having 87% of the data was allocated for training and another having 13% for testing. In the case of Linear models, data were subjected to feature scaling or normalization, while it was deemed unnecessary for nonlinear models. Also, Features like Cr and Ni having a high number of outliers were removed from the analysis while using Linear models in order to improve their performance [13]. Multiple regression, multiple regression with L1 regularization, multiple regression with L2 regularization, decision tree, and random forest models were established using the training data. The models' performance was assessed using R2 score, mean absolute percentage error (MAPE), and mean absolute error (MAE) metrics, which were later compared and have been summarized in Table 1.

2.1 Artificial Intelligence Models:

This study explores various machine learning models to estimate the yield strength of API steels on the basis of their chemical compositions and mechanical parameters. The aim is to achieve accurate prediction by implementing and evaluating several regression-based models, including random forest, decision tree, and multiple regression. This section provides a detailed overview of each of these models, including their underlying principles and performance metrics.

2.1.1. Multiple Regression:

Multiple regression is an analytical method used to analyze the relationship between multiple independent variables and an outcome variable. It helps to estimate the outcome variable based on the values of multiple independent features [14]. In multiple regression, a regression equation is created that estimates the value of the outcome variable based on the values of the independent variables.

The expression of multiple regression is given as:

$$\hat{y} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

here, $\hat{y}$ = Dependent Variable (Yield Strength(min)).
x_1 to x_n = independent variables (Chemical composition and mechanical properties)
E = Irreducible/ Prediction error

Hence the equation of the predicted yield strength can be given as:

$$Predicted\ yield\ strength = \alpha_0 + \alpha_1(C) + \alpha_2(Mn) + \alpha_3(Si) + \alpha_4(Mo) + \alpha_5(Cu) + \alpha_6(V) + \alpha_7(N) + \alpha_8(S) + \alpha_9(P) + \alpha_{10}(Fe) + \alpha_{11}\big(Hardness(HB,\ min)\big) + \alpha_{12}\big(Hardness(HB,\ max)\big) + E \quad (1)$$

where, α_0= y-intercept
α_n = Slope of the line, where n = 1,2, 3...

$$\alpha_0 = \frac{(\sum y)(\sum x^2)-(\Sigma x)(\Sigma xy)}{n(\sum x^2)-(\sum x)^2} \tag{1a}$$

$$\alpha_n = \frac{n(\sum xy)-(\sum x)(\Sigma y)}{n(\sum x^2)-(\sum x)^2} \tag{1b}$$

The above equations are solutions to the cost function of Multiple regression using the Ordinary Least Square method [15].

The cost function for Multiple regression:

$$\text{Minimize } \theta = \mathit{Minimize} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i = actual values and $\hat{y}_i$ = predicted values

As per our case, it can be written as:

θ = Summation of ((Actual Yield Strength – Predicted Yield strength)^2) for all records

However, in some cases, multiple regression models can be prone to overfitting, especially when there are many independent variables. This can result in models that are overly complex and have poor predictive power. Regularization is a method that can help in the prevention of overfitting by adding a penalty term to the regression equation [16].

L1 and L2 regularization are two commonly used techniques for regularized multiple regression. In the case of L1 regularization, a penalty term is added to the cost function that is proportional to the absolute value of the coefficients. This technique is also known as Lasso regression. On the other hand, in the case of L2 regularization, also called Ridge regression, a penalty term is added which is proportional to the square of the coefficients [17].

The cost function in case of L1 regularization:

$$\theta = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * (Sum\ of\ absolute\ values\ of\ coefficients)$$

$$\theta = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * \sum_{i=1}^{n}|\alpha_i| \tag{2a}$$

The cost function in case of L2 regularization:

$$\theta = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * (Sum\ of\ square\ of\ coefficients)$$

$$\theta = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda * \sum_{i=1}^{n}(\alpha_i)^2 \tag{2b}$$

The penalty term (λ) in regularization helps to control the size of the coefficients in the regression equation, preventing them from becoming too large and reducing the impact of irrelevant or redundant independent variables. By doing so, it helps improve the model's accuracy and stability [18].

Regularized multiple regression techniques like L1 and L2 regularization have proven to be effective in improving the multiple regression model's performance, particularly when dealing with large datasets with many independent variables.

2.1.2. Decision Tree:

Decision tree is a popular machine learning technique used for both regression and classification tasks. The basic idea behind this algorithm is to recursively partition the input data into smaller subsets, on the basis of the values of the input features, until each subset contains only one type of output (in classification) or a single value (in regression). Each partition is made using a decision rule that splits the data based on the value of a single feature [19].

The CART (Classification and Regression Trees) algorithm is a popular algorithm used to construct decision trees. The CART algorithm works by recursively splitting the data into two subsets, where each subset is split along a single feature. The feature and the splitting threshold are chosen to maximize the increase in homogeneity or reduction in impurity in the resulting subsets. The impurity of a subset is calculated using a measure of the variability of the output values in that subset. For classification tasks, commonly used impurity measures are entropy and Gini impurity, whereas for regression tasks, mean squared error (MSE) is typically used [20]. Equation 4, given below shows the cost function that the CART algorithm uses to minimize, to find the best split of a given node.

$$J = \frac{n_{left}}{n} MSE_{left} + \frac{n_{right}}{n} MSE_{right} \tag{3}$$

where,

$$MSE_{node} = \frac{\sum_{i \in node}\left(\hat{y} - y^{(i)}\right)^2}{n}$$

And,

$$\hat{y} = \frac{1}{n} \sum_{i \in node} y^{(i)}$$

In the above equations:

n = Records in the dataset
n_{left} = Records in the left subtree
n_{right} = Records in the right subtree

The CART algorithm begins with a single root node that represents the complete dataset. At each level of the tree, the algorithm chooses the feature and the splitting threshold that minimize the impurity or increase the homogeneity of the resulting subsets. The algorithm continues to split the data into smaller subsets until it reaches a stopping criterion, such as a minimum number of samples in each leaf node or a maximum tree depth [21].

Once the tree is constructed, it becomes feasible to make predictions on new data by traversing the tree from the root node to a leaf node. At each node, the decision rule corresponding to the splitting feature is applied, and the algorithm follows the appropriate branch based on the value of the feature in the input data. The prediction is then made based on the output value associated with the leaf node. Figure 2 given below, shows the visualization of decision tree generated by training the decision tree model on training data.

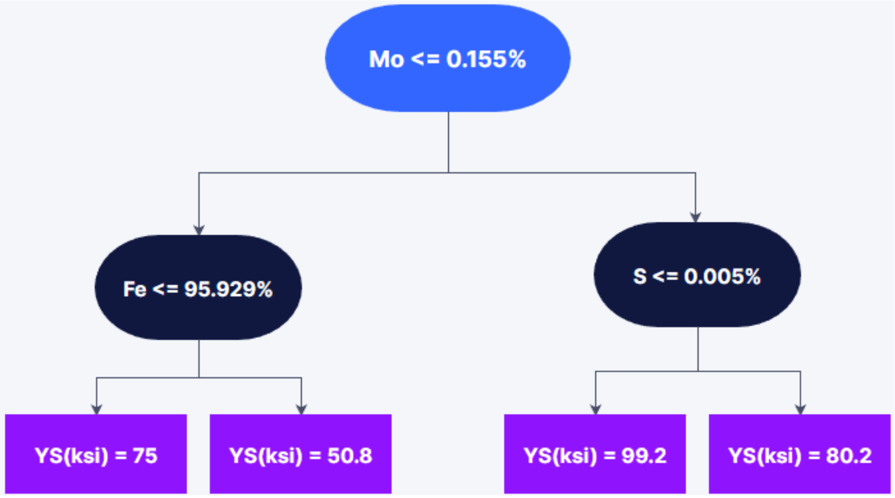

Figure 2: Decision Tree for Yield strength

2.1.3. Random Forest:

Random Forest is a widely used algorithm in machine learning used for both regression and classification tasks. It leverages the power of collective intelligence that works by constructing multiple decision trees and averaging their predictions [22] as shown in Figure 3.

In Random Forest regression, each tree in the ensemble is constructed using a random subset of the training data and a random subset of the input features [23]. This helps to improve the model's performance by reducing the correlation between the trees. The prediction of the Random Forest model is then obtained by averaging the predictions of all the individual trees [24].

Random Forest has several advantages over other regression models, including its capability to handle large datasets with many input variables, its capability to detect and handle relationships that are non-linear between input features and the target feature, and its capability to handle outliers and missing data [25]. On the other hand, it is not interpretable [26] and requires more training time and memory [27] due to the number of decision trees being generated.

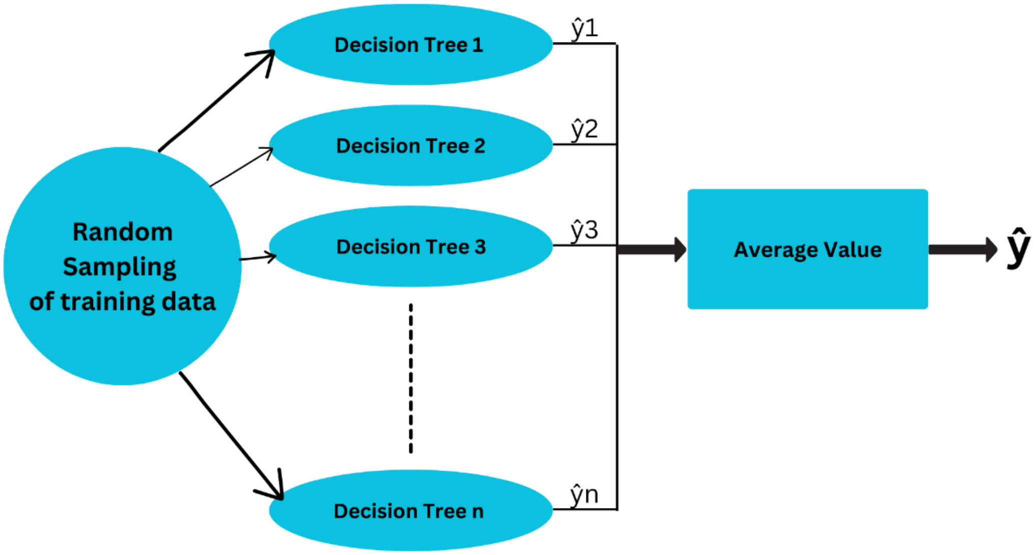

Figure 3: Random Forest flow diagram

2.2. Performance Metrics:

To judge the performance of our machine learning models, several evaluation criteria or performance metrics which are commonly used in regression problems were used. Each criterion provides a unique aspect of the model's performance, and together they give a detailed picture of how well the model is able to predict the yield strength of API steels.

Mean Absolute Error (MAE): MAE indicates that on average, how far the predicted values are from actual values, without taking their direction into consideration. It is calculated by taking the absolute difference between the real and estimated values and then averaging them. MAE has the same units as the variable being estimated. The lower the value of MAE, the better the model’s performance, since it means that the model's predictions are closer to the actual values [28].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{4}$$

Mean Absolute Percentage Error (MAPE): MAPE is a metric that quantifies the average percentage difference between the predicted and true values. To calculate the MAPE, the absolute percentage difference between the estimated and true values is obtained, and the resulting values are averaged. MAPE is expressed as a percentage. The lower the value of MAPE, the better the model’s performance, since it shows that the model's predictions are closer to the real values in percentage terms [29].

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right| \tag{5}$$

R-squared (R2) Score: The R2 score evaluates how much of the variability in the target variable can be accounted for by the independent variables incorporated in the model, ranging between 0 and 1. A higher R2 value indicates a better model’s performance. The calculation of R2 involves subtracting the ratio of the sum of squares of the residuals to the total sum of squares from 1. A R2 score of 1 denotes that the model fits the data perfectly, whereas a score of 0 indicates that the model fails to explain any variation in the target variable [30].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

These performance metrics were calculated for each of the machine learning models used in this study, which include Multiple Regression, Multiple Regression with L1 and L2 Regularization, Decision Tree, and Random Forest. The results are summarized in Table 1 and shown in Figure 9.

3. Results and Discussions:

In this section, the results of the study are presented, aimed at predicting the yield strength of API steels using various machine learning models. The performance of multiple regression, multiple regression with L1 and L2 regularization, decision tree, and random forest models was evaluated separately to predict yield strength. The test data was used to calculate the mean absolute percentage error (MAPE), mean absolute error (MAE), and R-squared (R2) score for each model. The following subsections present the results obtained to predict the yield strength of the API steels along with the interpretation of those results using feature importance.

Feature importance is a concept in machine learning that helps to identify the most important variables or features that are contributing the most to the model's prediction [31]. The importance of features is calculated based on the effect of that feature on a dependent variable. The higher the importance of a feature, the more significant its impact on the model's prediction accuracy [32]. Feature importance helps in identifying the relevant features that are driving the predictions, which can be useful to comprehend the existing relationships and making better decisions [33].

3.1. Prediction of AI models:

Based on the results obtained, the multiple regression model had a high MAE of 50.047 and a MAPE of 52.73%, indicating a poor performance in predicting yield strength. The Lasso regression model performed better than the multiple regression model with a MAE of 13.806 and a MAPE of 14.66%. The Ridge regression model also showed good performance with a MAE of 10.717 and a MAPE of 12.06%. The decision tree model had a MAE of 12.955 and a MAPE of 14.36%, indicating a slightly poorer performance than the Ridge regression model. Finally, the random forest model gave the best performance, with a MAE of 3.297 and a MAPE of 3.73%, and a R2 score of 0.9524.

In summary, the results imply that the random forest model is the most fruitful in predicting the yield strength of API steels, followed by Ridge regression, decision tree, lasso regression, and multiple regression.

3.1.1. Multiple regression:

In the multiple regression model, the feature importance for the prediction of yield strength can be interpreted based on the magnitude of the corresponding coefficients. Among the input variables, the ones with the highest magnitude coefficients have the most significant impact on the predicted yield strength. The coefficients show how the change in the independent variables will impact the dependent variable. For instance, As shown in Figure 4, a one-unit increase in Mn and Si will lead to a decrease in yield strength by 61.02 and 7.33 ksi respectively. Similarly, a one-unit increase in P and Mo will increase the yield strength by 56.25 and 33.30 ksi respectively. On the other hand, Fe and Hardness(HB, max) have a negative impact on yield strength, where a one-unit increase in these variables results in a decrease of 18.15 ksi and 17.56 ksi respectively.

The performance of the multiple regression model for yield strength prediction was evaluated using several metrics, including MAE, MAPE, and R2 score. The MAE of the model was found to be 50.047, indicating an average error of 50.047 ksi in the predicted yield strength values. The MAPE was 52.73%, suggesting that the model predictions were, on average, off by 52.73% of the actual yield strength values. The R2 score was -8.276, indicating that the model has poor predictive power and explains very little of the variance in the yield strength data. Overall, the model's performance needs to be improved to make accurate predictions of the yield strength of API steels.

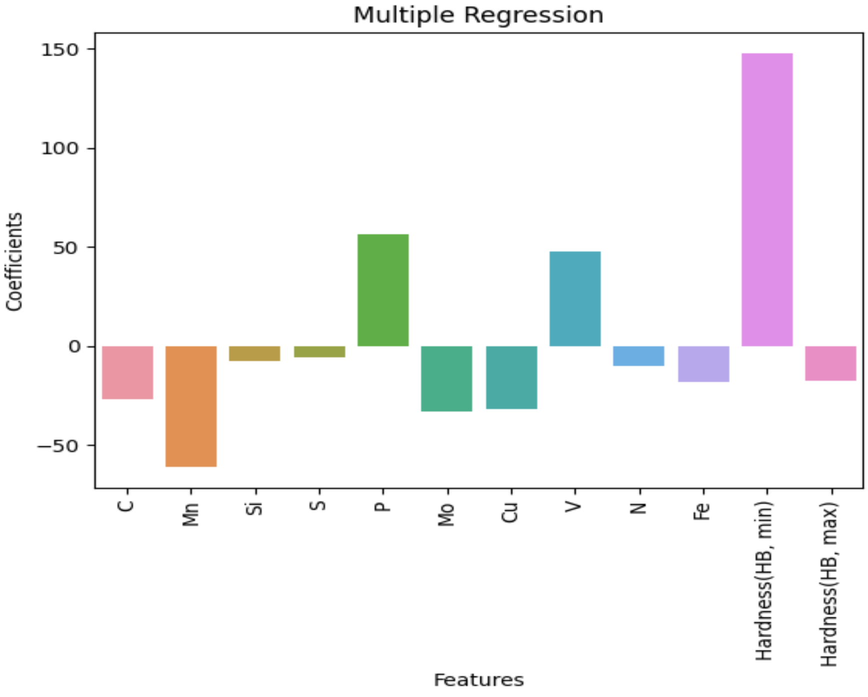

Figure 4: Feature importance w.r.t Multiple Regression Model

3.1.2. Multiple Regression with L1 regularization:

According to multiple regression with L1 regularization or Lasso regression, the important features for predicting yield strength having positive coefficients are the Hardness(HB, min) and Hardness(HB, max), i.e., one unit increase in their values would increase the yield strength by 1.8-65.43 ksi whereas the increment of one unit in the values of Mn, Cu, and S having negative coefficients would decrease the value of yield strength by 3.66-6.98 ksi, while other variables have coefficients close to 0, which means they have little or no influence on yield strength. Figure 5, suggests that the hardness, manganese, and copper content are important factors that affect the yield strength of API steels.

The evaluation of the model using the mean absolute percentage error (MAPE) and mean absolute error (MAE) shows a significant improvement in performance over the unregularized model. The MAE decreased from 50.047 to 13.806, while the MAPE decreased from 52.73% to 14.66%. The R2 score also improved from -8.276 to 0.1622. These results suggest that the L1 regularization technique has effectively reduced overfitting in the model and improved its predictive accuracy for yield strength prediction.

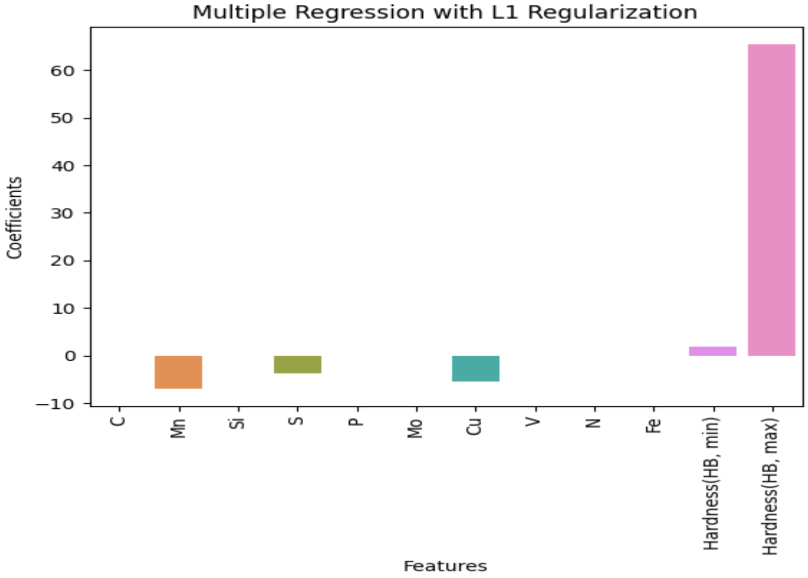

Figure 5: Feature importance w.r.t Lasso regression

3.1.3. Multiple Regression with L2 regularization:

In multiple regression with L2 regularization, also known as Ridge regression, as shown in Figure 6, the most important features having a positive impact on yield strength according to their coefficients are the Hardness(HB, max), Hardness(HB, min), C, Mo, N, and V ,i.e., one unit change in their value would increase yield strength by 1.8-23.6 ksi. On the other hand, Mn, Si, S, P, Cu, and Fe have negative coefficients, indicating that an increment in their values leads to a decrement in the predicted variable in the range of 1.03-11.33 ksi.

It has also been observed that the performance of the multiple regression model improved significantly after applying L2 regularization with a low MAE of 10.717, low MAPE of 12.06%, and higher R2 score of 0.6217.

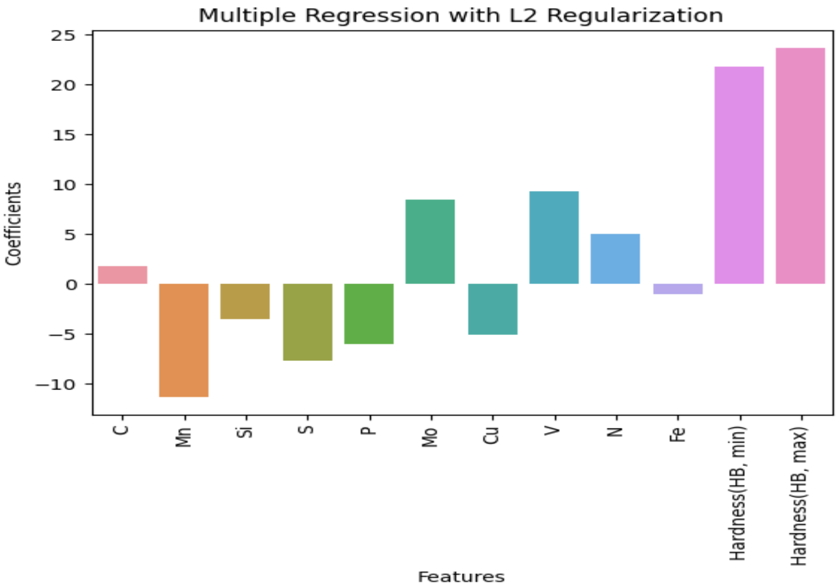

Figure 6: Feature importance w.r.t Ridge regression

3.1.4. Decision Tree:

In the decision tree model, the feature importance is calculated based on how much the feature contributes to the overall decision-making process of the model. The importance score of a feature is calculated by summing up the reduction in impurity [32] gained by splitting on that feature across all nodes in the tree. In this case, as shown in Figure 7, the feature importance values indicate that the 'Mo' is the most important feature for predicting yield strength, followed by the 'S' and 'Fe' features.

In terms of performance evaluation, the decision tree model achieves a mean absolute percentage error (MAPE) of 14.36% and a mean absolute error (MAE) of 12.955 which suggests that the model is capable of predicting the yield strength of API steel with reasonable accuracy. Additionally, the R2 score of 0.4813 indicates that the model explains a moderate proportion of the total variance in the yield strength of the API steel. Overall, the decision tree model provides good results, but its performance is not better than the Ridge regression.

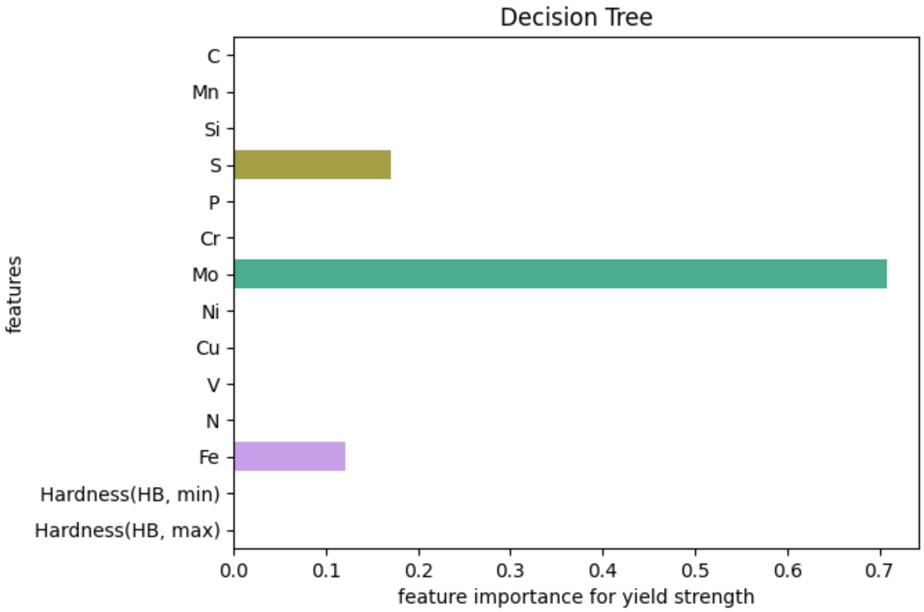

Figure 7: Feature importance w.r.t Decision Tree

3.1.5. Random Forest:

The feature importance for the Random Forest regression model indicates the relative importance of each feature in predicting the target variable. The feature importance is obtained based on the reduction in impurity (measured as mean squared error) achieved by splitting on a given feature over all trees in the forest. Specifically, for each tree, the decrease in impurity (impurity before split minus impurity after split) is computed for each feature that was used in the tree and then averaged over all trees in the forest [34].

Features that result in a high decrease in impurity when used for splitting tend to be more important, as they have a greater impact on reducing the error of the forest. The resulting feature importance values are normalized, to sum up to 1, so they can be interpreted as relative importance scores. In the case of yield strength, P is the most important, followed by Hardness (HB, min), and S, which has been shown in Figure 8.

In terms of model performance evaluation, the Random Forest regression model has a lower MAE and MAPE as 3.29 and 3.73% respectively, than the other models for yield strength, indicating that it is more accurate in predicting yield strength. The R2 score is 0.952 which is also higher, indicating that the Random Forest model explains more of the variance in yield strength. Overall, the Random Forest regression model showed promising results to estimate yield strength, with P, S, and Hardness (HB, min) being the most important features.

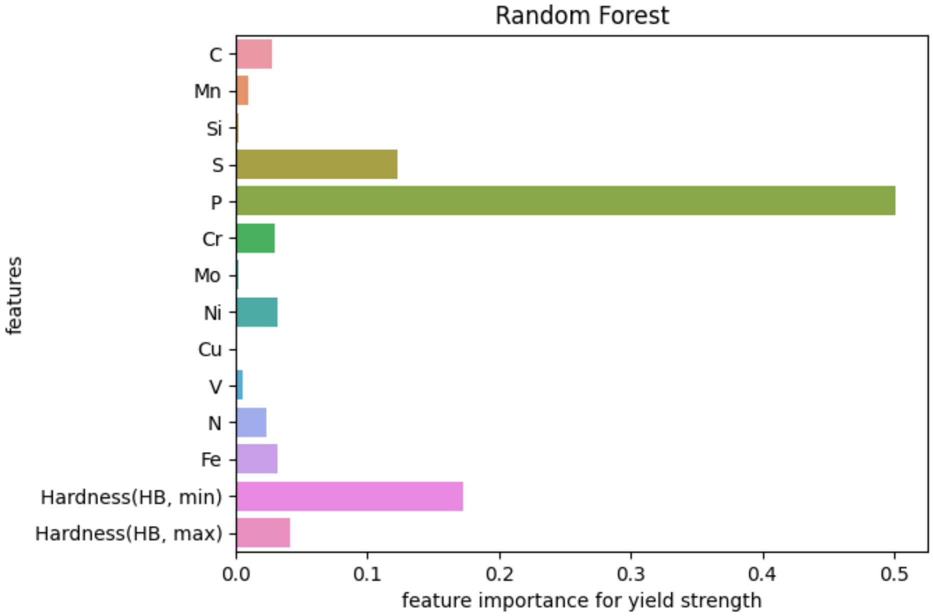

Figure 8: Feature importance w.r.t Random Forest model

Table 1: Models' Performance

Models	**Yield Strength**		
	MAE	**MAPE**	**R2 Score**
Multiple regression	50.047	52.73	-8.276
Lasso Regression	13.806	14.66	0.1622
Ridge Regression	10.717	12.06	0.6217
Decision Tree	12.955	14.36	0.4813
Random Forest	3.297	3.73	0.9524

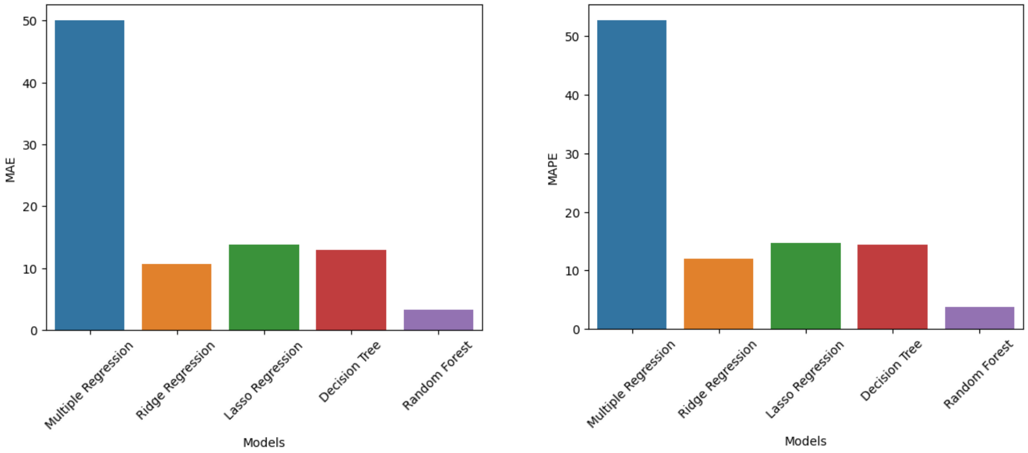

Figure 9: Visualization of models' performance

4. Conclusions:

In conclusion, the analysis conducted in this study has demonstrated that:

- The models developed in this study were implemented successfully and provided valuable insights into predicting the yield strength of API steels using machine learning techniques.
- The non-linear model (random forest) outperformed linear models in predicting the mechanical properties of API steels, specifically Yield strength.
- For the prediction of yield strength, the Random Forest model gave the most promising results with a MAE of 3.29, a MAPE of 3.73, and a R2 score of 0.95.

References:

1. Ananta Nagu, G. & G Namboodhiri, T. K. Effect of heat treatments on the hydrogen embrittlement susceptibility of API X-65 grade line-pipe steel. *Bull. Mater. Sci* **26**, 435–439 (2003).
2. Das, A. K. The present and the future of line pipe steels for petroleum industry. *Center for Bioinformatics and Molecular Biostatistics* **25**, 14–19 (2010).
3. 5 Common Alloying Elements. https://www.metalsupermarkets.com/5-common-alloying-elements/.
4. W.E. White & G. I. Ogundele. Influences of Dissolved Hydrocarbon Gases and Variable Water Chemistries on Corrosion of an API-L80 Steel. *CORROSION* **43**, 665–673 (1987).
5. Elgaddafi, R., Ahmed, R. & Shah, S. Modeling and experimental studies on CO2-H2S corrosion of API carbon steels under high-pressure. *J Pet Sci Eng* **156**, 682–696 (2017).
6. A brief introduction of API 5L, API 5B, API 5CT and API 5D. https://www.worldironsteel.com/news/a-brief-introduction-of-api-5l-api-5b-api-5c-12677821.html.
7. In depth guide to machine learning. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.
8. Meng, L. *et al.* Machine Learning in Additive Manufacturing: A Review. *JOM* **72**, 2363–2377 (2020).
9. Gan, L., Wu, H. & Zhong, Z. Fatigue life prediction considering mean stress effect based on random forests and kernel extreme learning machine. *Int J Fatigue* **158**, (2022).
10. Nasiri, S. & Khosravani, M. R. Machine learning in predicting mechanical behavior of additively manufactured parts. *Journal of Materials Research and Technology* **14**, 1137–1153 (2021).
11. Sah, A. K., Agilan, M., Dineshraj, S., Rahul, M. R. & Govind, B. Machine learning-enabled prediction of density and defects in additively manufactured Inconel 718 alloy. *Mater Today Commun* **30**, (2022).
12. Carvalho, T. P. *et al.* A systematic literature review of machine learning methods applied to predictive maintenance. *Comput Ind Eng* **137**, (2019).
13. Moeedlodhi. How Outliers Can Pose a Problem in Linear Regression. https://medium.com/swlh/how-outliers-can-pose-a-problem-in-linear-regression-1431c50a8e0 (2020).
14. Wagner, M. M., Moore, A. W. & Aryel, R. M. Combining Multiple Signals for Biosurveillance. in *Handbook of Biosurveillance* 235–242 (Academic Press, 2006).
15. Induraj. How to derive B0 and B1 in Linear Regression- Part2. https://induraj2020.medium.com/how-to-derive-b0-and-b1-in-linear-regression-4d4806b231fb (2020).
16. Ying, X. An Overview of Overfitting and its Solutions. *J Phys Conf Ser* **1168**, (2019).
17. L.E. Melkumova & S.Ya. Shatskikh. Comparing Ridge and LASSO estimators for data analysis. *Procedia Eng* **201**, 746–755 (2017).
18. How do you explain the impact of regularization on the bias-variance trade-off in linear regression?

https://www.linkedin.com/advice/0/how-do-you-explain-impact-regularization-bias-variance

19. S. B. Kotsiantis. Decision trees: A recent overview. *Artif Intell Rev* **39**, 261–283 (2013).
20. Roman Timofeev. Classification and Regression Trees (CART) Theory and Applications. (Humboldt University, 2004).
21. Max Bramer. Avoiding Overfitting of Decision Trees. in *Principles of Data Mining* 119–134 (Springer, 2007).
22. Breiman, L. Random Forests. **45**, 5–32 (2001).
23. Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. (2019).
24. Mark R. Segal. Machine Learning Benchmarks and Random Forest Regression. *Center for Bioinformatics and Molecular Biostatistics* 1–14 (2004).
25. Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing* **67**, 93–104 (2012).
26. Ziegler, A. & König, I. R. Mining data with random forests: Current options for real-world applications. *Wiley Interdiscip Rev Data Min Knowl Discov* **4**, 55–63 (2014).
27. Santur, Y., Karaköse, M. & Akın, E. Random Forest Based Diagnosis Approach for Rail Fault Inspection in Railways. *National Conference on Electrical, Electronics and Biomedical Engineering (ELECO)* 714–719 (2016).
28. Mean absolute error. https://en.wikipedia.org/wiki/Mean_absolute_error.
29. Mean absolute percentage error. https://en.wikipedia.org/wiki/Mean_absolute_percentage_error.
30. Coefficient of determination. https://en.wikipedia.org/wiki/Coefficient_of_determination.
31. Mirka Saarela & Susanne Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Appl Sci* **3**, (2021).
32. Jason Brownlee. How to calculate feature importance with python. https://machinelearningmastery.com/calculate-feature-importance-with-python/ (2016).
33. Aaron Fisher, Cynthia Rudin & Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* **20**, 1–81 (2019).
34. Christine Dewi & Rung-Ching Chen. Random forest and support vector machine on features selection for regression analysis. *International Journal of Innovative Computing, Information and Control* **15**, 2027–2037 (2019).