# Machine Learning Approaches for Predicting Structure-Property Relationship in Industrial Steels

[Saikiran Anugam[1], Hariharan Annadurai[2]]

[[1]Department of Engineering Data science, [2]Department of Chemical and Biomolecular Engineering], University of Houston

## Abstract

In this study, we explore the relationship between chemical compositions and mechanical properties specifically Tensile Strength(TS) of industrially produced steels using a comprehensive dataset. We apply various machine learning techniques such as Linear Regression, Random Forest, Support Vector Regression (SVR), K-Nearest Neighbours (KNN) regression to map the features to properties of steel. We evaluate model performance using Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) with extensive hyperparameter tuning. Feature selection is utilized to identify the most influential chemical components impacting mechanical properties. Our findings elucidate the critical chemical composition of elements essential for required tensile, aiding in identifying brittleness and indicating the elastic stress a material can withstand. This research enhances the current understanding of materials science, providing insights into steel properties with practical implications for optimizing steel manufacturing processes and improving material performance. The use of advanced models and robust error analysis strengthens the applicability and reliability of our findings.

**Keywords:** Mean Square Error(MAE), Random Forest Regressor, Support Vector Regressor(SVR),Tensile Strength

## Background

Steel is a versatile and widely used material known for its strength, durability, and adaptability. It is primarily composed of iron and carbon, with properties that can be significantly altered by adjusting the carbon content and adding various alloying elements such as manganese, chromium, nickel, and titanium.

It is broadly classified based on composition and use, including low, medium, and high carbon steel, as well as alloy steels like stainless steel and tool steel. Its diverse properties make it essential in industries such as construction, automotive, machinery, energy, and household goods.

Understanding mechanical properties like tensile strength, yield strength, and elongation is crucial for steel's applications. By optimizing chemical composition and processing methods, engineers can design steels with tailored properties for specific uses, ensuring safety, efficiency, and longevity.

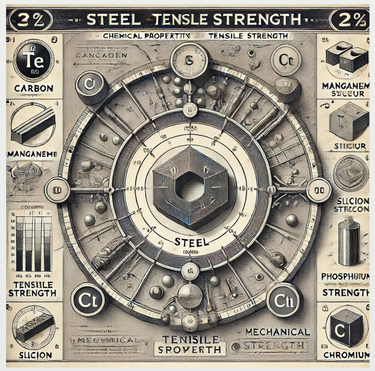


*Fig.1: Chemical Properties & Tensile Strength*

## Methods

### Feature Selection Techniques

#### Univariate Feature Selection (ANOVA F-test)

- We used the SelectKBest method from sklearn feature selection to apply the ANOVA F-test, which compares the variance between chemical properties and Tensile Strength.
- S (Sulphur), Ca (Calcium), and Nb (Niobium) are the most significant predictors of tensile strength.
- Elements like W (Tungsten) and Sb (Antimony) have much lower F-scores, indicating a lesser impact on tensile strength.

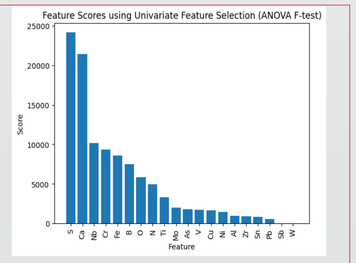


**Fig.2: Feature Selection using ANOVA F-test**

#### Feature Selection using Pearson Correlation Coefficient

**Top Features**:

- Ti (Titanium): Highest MI score (0.84), indicating it has the most significant impact on Tensile Strength.
- Fe (Iron): Second highest MI score (0.61), suggesting a strong relationship with the target variable.
- Nb (Niobium): Also shows a high MI score (0.51), indicating its importance.

**Less Influential Features:**

- Sn (Tin): Low MI score (0.07), suggesting it has a minimal impact on the target variable.
- W (Tungsten): MI score of 0, indicating no dependency with the target variable.

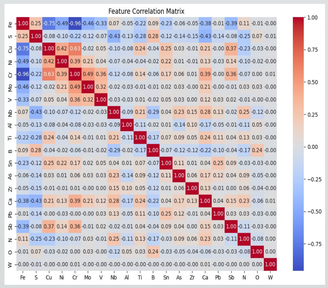


Fig.3:Heatmap visualizing correlation matrix

#### Feature Selection using Mutual Information

- Features like Fe and Cr have a strong negative correlation (-0.96), indicating that as the concentration of Fe increases, the concentration of Cr decreases.
- Similarly, S and Ca have a strong negative correlation (-0.86).
- Ti and Zr have a strong positive correlation (0.80), suggesting that these elements tend to increase together.

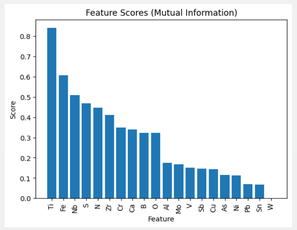


*Fig.4: Feature scores using Mutual information*

#### Recursive Feature Elimination using Random Forest Regressor

**Top Features:**

- S (Sulphur): Highest importance score, indicating it has the most significant impact on Tensile Strength. O (Oxygen) .
- Ti (Titanium): Also show high importance scores, suggesting strong relevance.

**Less Influential Features:**

Cu (Copper) and Al (Aluminum): Lower importance scores, indicating less impact on the target variable compared to other selected features.

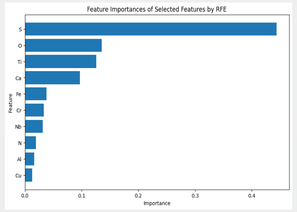


*Fig.5: Feature scores using Random Forest Regressor*

#### Dimensionality Reduction using Principal Component Analysis

- In this case, 12 principal components explain around 80% of the variance, indicating that a significant reduction in dimensionality can be achieved while retaining most of the information.

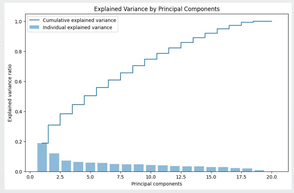


*Fig.6: Explained variance vs. Principal components*

- **Analysis of the Plot:**

**Strong Positive Contributors:**

- Cr (Chromium): Highest positive loading score, indicating a strong positive contribution to the first principal component.
- Ni (Nickel) and Mo (Molybdenum): Also have high positive loading scores.

**Strong Negative Contributors:**

- Fe (Iron): Highest negative loading score, indicating a strong negative contribution to the first principal component.
- S (Sulphur): Also has a significant negative loading score.

**Moderate Contributors:**

- Ca (Calcium) and Pb (Lead): Moderate positive loading scores, contributing positively but not as strongly as Cr or Ni.
- Nb (Niobium) and Al (Aluminum): Moderate negative loading scores.

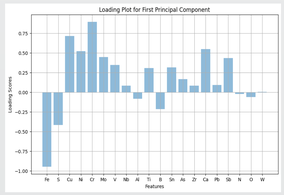


*Fig.7: Loading Plot for first principal component*

### Machine learning models

**Linear Regression**.

- RFE shows the least performance improvement, indicating it may not be the best choice for this specific model

**KNN Regression**

- Best Hyperparameters using Grid Search CV metric: Manhattan, nearest neighbors: 9, weights: distance

**Random Forest**

- Best hyperparameters found during the randomized search are: number_of_estimators:200, min_samples_split: 5, min_samples_leaf: 1, max_depth: None

**Support Vector Regressor**

- The best hyperparameters for the SVR model using Grid search were found to be using the RBF kernel with C of 100, epsilon of 1, and gamma set to 'scale'.

| Model | Feature Selection Method | Testing MAE | Testing MSE | Testing RMSE |
|---|---|---|---|---|
| SVR | All Features | 27.43 | 1964.05 | 44.32 |
| SVR | Correlation Features | 27.32 | 1895.69 | 43.54 |
| SVR | MI Features | 26.96 | 1848.60 | 43.00 |
| SVR | RFE Features | 28.75 | 2006.76 | 44.80 |
| SVR | PCA | 31.41 | 2542.75 | 50.43 |
| KNN Regressor | All Features | 17.57 | 736.26 | 27.13 |
| KNN Regressor | Correlation Features | 17.54 | 726.54 | 26.95 |
| KNN Regressor | MI Features | 17.57 | 730.04 | 27.02 |
| KNN Regressor | RFE Features | 17.77 | 756.44 | 27.50 |
| KNN Regressor | PCA | 25.46 | 2257.74 | 47.52 |
| Random Forest | All Features | 16.93 | 655.05 | 25.59 |
| Random Forest | Correlation Features | 17.10 | 669.20 | 25.87 |
| Random Forest | MI Features | 17.01 | 661.05 | 25.71 |
| Random Forest | RFE Features | 17.19 | 673.38 | 25.95 |
| Random Forest | PCA | 33.86 | 3179.54 | 56.39 |
| Linear Regression | All Features | 50.41 | 5427.98 | 73.67 |
| Linear Regression | Correlation Features | 52.35 | 5883.25 | 76.70 |
| Linear Regression | MI Features | 52.29 | 5878.36 | 76.67 |
| Linear Regression | RFE Features | 53.85 | 6255.19 | 79.09 |

## Results

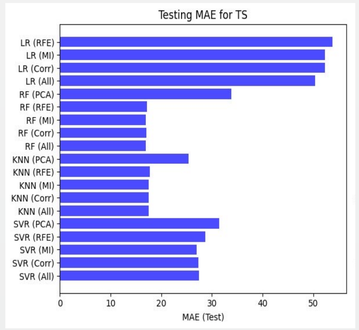


*Fig.8:Testing MAE*

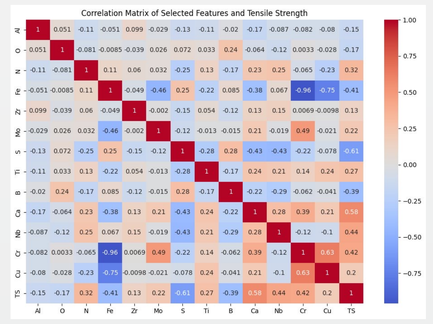


*Fig.9: Correlation matrix of selected features and tensile strength*

**Frequency of Features Selected by Different Methods**

**Correlation-based Selection:**
Selected features: ['Fe', 'S', 'Cr', 'Mo', 'Nb', 'Ti', 'B', 'Ca', 'N', 'O'].

**Mutual Information (MI):**
Selected features: ['Ti', 'Fe', 'Nb', 'S', 'N', 'Zr', 'Cr', 'Ca', 'O', 'B'].

**Recursive Feature Elimination (RFE):**
Selected features: ['Fe', 'S', 'Cu', 'Cr', 'Nb', 'Al', 'Ti', 'Ca', 'N', 'O'].

**Analysis of the Output:**

**Highly Frequent Features:**
Fe (Iron), S (Sulphur), Cr (Chromium), Nb (Niobium), Ti (Titanium), Ca (Calcium), N (Nitrogen), O (Oxygen):
These features were selected by all three methods, indicating their high importance and relevance.
Frequency: 3

**Moderately Frequent Features**:
B (Boron):
Selected by two methods.
Frequency: 2

**Less Frequent Features**:
Mo (Molybdenum), Zr (Zirconium), Cu (Copper), Al (Aluminum):
Selected by only one method.
Frequency: 1

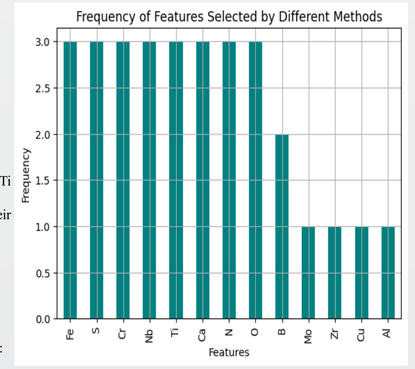


*Fig.10: Frequency of features selected by different methods*

## Conclusion

**Model Accuracy:** The Random Forest model, using features selected through Mutual Information (MI), provides the most accurate predictions for tensile strength in industrial steels.

**Feature Integration**: Combining common features from all selection methods enhances the model's robustness and accuracy**.**

**Key Elements**: Titanium, chromium, and niobium are crucial for improving tensile strength by controlling grain size, forming precipitates, and through precipitation hardening.

**Role of Sulfur:** Sulfur's traditional role in deoxidation, alongside its interaction with transition metals, is particularly significant in this study.

**Validation**: The findings align with metallurgical principles, confirming the importance of specific elements in enhancing mechanical properties, making the results applicable to steel production and improvement

## Future work

**Extended Analysis:** The study can be extended to analyze other mechanical properties of steel, such as yield strength and elongation**.**

**Feature Selection:** Applying similar feature selection techniques and machine learning models can identify the key chemical elements influencing these properties.

**Comprehensive Comparison:** Comparing the important features across tensile strength, yield strength, and elongation will provide a thorough understanding of chemical influences on mechanical properties.

**Stress-Strain Insights:** Analyzing the stress-strain ratio will offer deeper insights into the material's behavior under stress, aiding in the development of high-performance steels

## References

- Guo, S., Yu, J., Liu, X., Wang, C., & Jiang, Q. (2019). A predicting model for properties of steel using the industrial big data based on machine learning. Computational Materials Science, 160, 95–104. https://doi.org/10.1016/j.commatsci.2018.12.056
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- Campbell, F. C. (2008). Elements of metallurgy and Engineering Alloys. ASM International