KADIR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
PROGRAM OF MANAGEMENT INFORMATION SYSTEM

PREDICTIVE MODELLING OF MECHANICAL PROPERTIES IN FLAT STEEL MANUFACTURING: A COMPARATIVE ANALYSIS OF FEATURE SELECTION METHODS OF MECHANICAL PROPERTIES FOR COLD ROLLED PRODUCTS

DİDEM BAKİLER İLME

MASTER OF SCIENCE THESIS

ISTANBUL, JULY 2024

PREDICTIVE MODELLING OF MECHANICAL PROPERTIES IN FLAT STEEL MANUFACTURING: A COMPARATIVE ANALYSIS OF FEATURE SELECTION METHODS OF MECHANICAL PROPERTIES FOR COLD ROLLED PRODUCTS

DİDEM BAKİLER İLME
ADVISOR: DOÇ. DR. EMRULLAH FATİH YETKİN

A thesis submitted to
the School of Graduate Studies of Kadir Has University
in partial fulfillment of the requirements for the degree of
Master of Science in
Management Information Systems

Istanbul, June 2024

APPROVAL

This thesis titled PREDICTIVE MODELLING OF MECHANICAL PROPERTIES IN FLAT STELL MANUFACTURING: A COMPARATIVE ANALYSIS OF FEATURE SELECTION METHODS OF MECHANICAL PROPERTIES FOR COLD ROLLED PRODUCTS submitted by DİDEM BAKİLER İLME, in partial fulfillment of the requirements for the degree of Master of Science in Management Information Systems is approved by

Assoc. Prof. Dr. E. Fatih YETKİN (Advisor) ……..
Kadir Has University

Asst. Prof. Dr. Tuğçe BALLI ……..
Kadir Has University

Asst. Prof. Dr. Şenol PİŞKİN ……..
İstinye University

I confirm that the signatures above belong to the aforementioned faculty members.

_____

Prof. Dr. Mehmet Timur AYDEMİR
Director of the School of Graduate Studies
Date of Approval: 24.07.2024

DECLARATION ON RESEARCH ETHICS AND PUBLISHING METHODS

I, DİDEM BAKİLER İLME, hereby declare

• This Master of Science Thesis that I have submitted is entirely my own work, and I have cited and referenced all material and results that are not my own by the rules.

• This Master of Science Thesis does not contain any material from any research submitted or accepted to obtain a degree or diploma at another educational institution.

• and that I commit and undertake to follow the "Kadir Has University Academic Codes and Conduct" prepared by the "Higher Education Council Codes of Conduct." In addition, I acknowledge that any claim of irregularity that may arise about this work will result in disciplinary action by university legislation.

DİDEM BAKİLER İLME

_____

Date: 24/07/24

To My Dearest Family, Mesut & Demirhan & Gürhan…

ACKNOWLEDGEMENT

I would like to express my profound gratitude to my supervisor, Assoc. Prof. Dr. E. Fatih Yetkin, for his exceptional guidance and patience throughout my thesis work. His expertise and insightful critiques have been invaluable.

I am also immensely thankful to the members of my thesis committee for their constructive feedback and encouragement.

My sincere appreciation extends to my colleagues, Merve Öper, Aslı Koca, Saygın Kacar, Duygu Horoz, and Asena Öztürk for their support throughout this journey.

I must also express my heartfelt thanks to my family. To my spouse, Mesut İlme, for his understanding and endless love; and to my children, Demirhan Çınar İlme and Mehmet Gürhan İlme, for their patience and joy that light up my life.

Thank you all for being my constant support and for believing in me.

PREDICTIVE MODELING OF MECHANICAL PROPERTIES IN FLAT STEEL MANUFACTURING: A COMPARATIVE ANALYSIS OF FEATURE SELECTION METHODS OF MECHANICAL PROPERTIES FOR COLD ROLLED PRODUCTS

ABSTRACT

Cold-rolled flat steel coil products play a critical role across various industries and possess multiple applications. The essential nature of these products primarily stems from their excellent mechanical properties. Cold-rolled flat steel coils are utilized as semi-products in the automotive, home appliance, radiator, construction, and packaging industries. Through the cold rolling process of flat steel, the parameters of cold rolling, batch annealing, and skin pass processes are optimized according to the application area's requirements, thus achieving the desired mechanical properties in compliance with the relevant standards.

This study uses machine learning algorithms to eliminate the need to take mechanical property test samples on the skin pass line in the flat steel industry. This approach will help reduce the scrap rate and will result in capacity gains in production. Within this study, machine learning models such as Linear Regression (LR), Support Vector Regressors (SVR), K-nearest neighbors (KNN), Random Forest (RF), XGBoost, and Decision Tree (DT) have been performed to predict yield and tensile strength of flat steel products. The performance of the models is improved by running eight different feature selection methods alongside hyperparameter tuning and cross-validation.

In evaluating yield strength, applying the XGBoost model across the complete dataset achieved a coefficient of determination R^2 of 93.8%. Focusing on data specific to the European Union (EU), the R^2 improved to 95%, indicating superior model performance. The KNN model yielded an R^2 of 91.9% for the Japanese dataset. Further refinement using the Least Absolute Shrinkage and Selection Operator (LASSO) method with the XGBoost model on the EU dataset elevated the R^2 to 95.4% upon incorporating the 23rd feature and %. Further refinement using the Sequential Forward Selection (SFS) method and the XGBoost model on the Japan dataset elevated the R^2 to 94.5% upon incorporating the 23rd feature.

Regarding tensile strength prediction, the XGBoost model demonstrated a robust performance across the entire dataset, attaining an R^2 of 90.8%. In contrast, the KNN

model, when applied to the EU data, reached an R^2 of 88.4%. Applying the same model to the Japanese data resulted in an R^2 of 89.6%. Employing a combination of Mutual Information, ANOVA F-Test, and recursive feature selection strategies on the full dataset improved the R^2 marginally to 91.2%. In the EU context, adapting the mutual feature selection method boosted the R^2 to 89.5% when using the KNN algorithm. Conversely, in the Japanese dataset, the KNN model's performance enhanced to an R^2 of 91.6% after applying the Sequential Forward Selection (SFS) method at the 22nd feature. The least effective approach across yield and tensile strength predictions was Linear Regression.

The outcome of this thesis provides compelling evidence that the application of machine learning algorithms, particularly the XGBoost model, significantly enhances the prediction accuracy of mechanical properties in cold-rolled flat steel coils, thereby potentially eliminating the need for physical test samples on the skin pass line and reducing both scrap rates and production costs across the flat steel industry.

Keywords: Machine Learning Models, Regression Methods, Feature Selection Methods, Material Science, Flat Steel

YASSI ÇELİK ÜRETİMİNDE MEKANİK ÖZELLİKLERİN TAHMİNSEL MODELLEMESİ:
SOĞUK HADDELENMIİŞ ÜRÜNLER İÇİN MEKANİK ÖZELLİKLERİN ÖZELLİK SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMALI ANALİZİ

ÖZET

Soğuk haddelenmiş yassı çelik rulo ürünler çeşitli sektörlerde kritik bir rol oynamaktadır ve çok çeşitli uygulamalara sahiptir. Bu ürünlerin esas önemi, mükemmel mekanik özelliklerinden kaynaklanmaktadır. Soğuk haddelenmiş yassı çelik bobinler, otomotiv, ev aletleri, radyatör, inşaat ve ambalaj endüstrilerinde yarı ürün olarak kullanılmaktadır. Yassı çeliklerin soğuk haddeleme sürecinde, soğuk haddeleme, tavlama ve son soğuk haddeleme süreçlerinin parametreleri, uygulama alanının gereksinimlerine göre optimize edilerek ilgili standartlara uygun olarak arzu edilen mekanik özellikler elde edilmektedir.

Bu çalışma, düz çelik endüstrisinde mekanik özellik test örneklerinin alınması gerekliliğini ortadan kaldırmak için makine öğrenimi algoritmalarını kullanmaktadır. Bu yaklaşım, hurda oranını azaltmaya ve üretimde kapasite kazançları sağlamaya yardımcı olacaktır. Bu çalışmada, düz çelik ürünlerin akma ve çekme mukavemetini tahmin etmek için Doğrusal Regresyon (LR), Destek Vektör Regresörleri (SVR), K-en yakın komşular (KNN), Rastgele Orman (RF), XGBoost ve Karar Ağacı (DT) gibi makine öğrenimi modelleri kullanılmıştır. Modellerin performansı, sekiz farklı özellik seçme yöntemi ile hiperparametre ayarlaması ve çapraz doğrulama yapılarak geliştirilmiştir.

Akma mukavemetinin değerlendirilmesinde, XGBoost modelinin tüm veri seti üzerinde uygulanması %93.8 oranında bir belirleme katsayısı (R^2) başarısı elde etmiştir. Avrupa Birliği (AB) verilerine özgü verilere odaklanıldığında, R^2 %95'e yükselmiştir. Japon veri seti için KNN modeli %91.9 R^2 elde etmiştir. Avrupa tedarikçileri veri setinde XGBoost modeli ile LASSO yönteminin birlikte kullanılması 23. özellik eklenerek R^2'yi %95.4'e çıkarmıştır. Japonya veri setinde Sıralı İleri Seçim (SFS) yöntemi ve XGBoost modeli ile yapılan daha ileri düzey bir rafinasyon, R^2'yi 23. özellik eklenerek %94.5'e çıkarmıştır.

Çekme mukavemeti tahmini konusunda, XGBoost modeli tüm veri setinde güçlü bir performans sergileyerek %90.8 R^2 elde etmiştir. Buna karşılık, Avrupa tedarikçileri verilerine uygulanan KNN modeli %88.4 R^2'ye ulaşmıştır. Aynı modelin Japon verilerine uygulanması %89.6 R^2 sonucunu vermiştir. Tüm veri setinde Karşılıklı Bilgi, ANOVA

F-Testi ve yinelemeli özellik seçim stratejilerinin bir kombinasyonu R^2'yi hafifçe %91.2'ye yükseltmiştir. Avrupa veri seti kullanıldığında, karşılıklı özellik seçim yönteminin uyarlanması KNN algoritması kullanılarak R^2'yi %89.5'e çıkarmıştır. Buna karşın, Japon veri setinde, KNN modelinin performansı Sıralı İleri Seçim (SFS) yöntemi uygulanarak 22. özellikte %91.6 R^2'ye çıkarılmıştır. Akma ve çekme mukavemeti tahminlerinde en az etkili yaklaşım Lineer Regresyon olmuştur.

Bu tezin sonucu, özellikle XGBoost modelinin kullanımının, soğuk haddelenmiş düz çelik bobinlerin mekanik özelliklerinin tahmin doğruluğunu önemli ölçüde artırarak, son haddeleme hattında fiziksel test örneklerine olan ihtiyacı potansiyel olarak ortadan kaldırabileceği ve yassı çelik endüstrisinde hem hurda oranlarını azaltabileceği hem de üretim maliyetlerini düşürebileceği yönünde ikna edici kanıtlar sunmaktadır.

Anahtar Sözcükler: Makine Öğrenimi Modelleri, Regresyon Yöntemleri, Özellik Seçim Yöntemleri, Malzeme Bilimi, Yassı Çelik

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Y ..... Dependent variable

β_0 ..... Y-intercept

$\beta 1, \beta 2, \ldots, \beta p$ ..... Independent variable coefficients

$X_1, X_2, \ldots X_p$ ..... Independent variables

ε_i ..... Error term

i ..... Observation index

n ..... Number of observations

z ..... Z-score (standardized value)

x ..... Attribute value (original data)

μ (mu) ..... Attribute mean (dataset).

σ (sigma) ..... Attribute standard deviation

x_{tr} ..... Feature training data.

x_{test} ..... Feature test data.

y_{tr} ..... Target training data.

y_{test} ..... Target test data.

LIST OF ACRONYMS AND ABBREVIATIONS

AI:	Artificial Intelligence
ACC:	Accelerated Cooling Control
ANN:	Artificial Neural Networks
CE:	Carbon Equivalent
CNN:	Convolutional Neural Network
DNN:	Deep Neural Network
EN:	European Norm
FH:	Full Hard
DL:	Deep Learning
DT:	Decision Tree
KNN:	K-nearest neighbors
LASSO:	Least Absolute Shrinkage and Selection Operator
LR:	Linear Regression
ML:	Machine Learning
MLP:	Multilayer perceptron
NLP:	Natural Language Processing
RF:	Random Forest
SFS:	Sequential Feature Selector
SVR:	Support Vector Regressors
XGBoost:	eXtreme Gradient Boosting

1. INTRODUCTION

Cold-rolled flat steels are commonly used in many industries like automotive sector, home appliances, and construction (Q Xie et al., n.d.), (Beranger, Henry, and Sanz 1994). These steels are compared for their advanced surface qualities, their precise dimensions and their enhanced mechanical properties. These properties make them good for high-performance applications. The mechanical properties of cold-rolled flat steel, especially yield and tensile strength, are very important for their suitability in different applications. Standards such as the European Norm (EN) and specific automotive standards ensure these materials meet strict requirements for formability and structural integrity (Reddy et al. 2020). Yield strength is the ability of a material to withstand deformation under tensile load. It marks the transition from elastic to plastic deformation. This property is very important for the design of components that must maintain their shape under stress. High yield strength is also significant for applications under significant forces, like buildings, bridges, and automotive components. Also, tensile strength is the maximum stress a material can endure while being stretched before fracturing. This indicates its ultimate strength and resistance to breaking. This property is vital in engineering and materials science as it defines the material's performance limits and safety margins (Beranger, Henry, and Sanz 1994)

Traditionally, the tensile and yield strengths of cold-rolled flat steels are measured by the tensile test method. For this test, the material is prepared and subjected to the tensile load until it fractures. The test provides valuable data on the material's mechanical properties but has several disadvantages. Sample preparation consumes time, process times increase, and causes higher costs. Additionally, because the test is destructive, the samples cannot be reused, resulting in material wastage. These limitations have challenges, especially in high-volume industrial environments where efficiency and cost-effectiveness are important (E. Dowling Norman, Siva Prasad Katakam, and R. Narayansamy 2013).

Machine learning methods are a good alternative to solve these problems. These algorithms can analyze the previous data from production processes to predict the

mechanical properties of materials, and it eliminates the need for destructive testing. The historical data is used such as chemical compositions, processing parameters, and mechanical test results, and then machine learning models could be trained to understand patterns and correlations that affect yield and tensile strength. Consequently, these models can predict the properties of new steel batches based on their production data.
In the final process of cold-rolled flat steel products, which is the skin pass line, a mechanical property sample is taken to meet customer demands. The mechanical property prediction model will provide two fundamental benefits: efficiency and cost reduction. Initially, during the sampling process, the last about eight meters of the coil are scraped, and a sample is taken across a width of 500 mm from the point where the line enters its regime. This process takes approximately 1 to 1.5 minutes. From this sample taken from the line, a mechanical property test sample is extracted, and a destructive mechanical property test is conducted. By employing machine learning algorithms to develop a predictive model for mechanical properties, the eight meters of scrap from each coil will no longer need to be removed. Additionally, the sampling process, which takes about 1 to 1.5 minutes per coil, will no longer be necessary, thus increasing line capacity. The capacity gain per coil, calculated at approximately 0.5 tons, is a significant improvement.

This study aims to predict the yield and tensile strength of coils in the flat steel industry by proposing a machine learning pipeline using six different models (Random Forest, Decision Tree, Support Vector Regression, Linear Regressor, K-nearest Neighbor, and extreme gradient boosting) incorporating 24 distinct production parameters as inputs. To enhance the predictive performance of these models for yield and tensile strength, seven different feature selection methods (Mutual Information. ANOVA F-Test, Recursive Feature Selection, Sequential Feature Selection, LASSO, Random Forest, XGBoost) are employed. These methods systematically rank the production parameters from most to least influential and are iteratively utilized within the models to refine their accuracy. Applying these feature selection techniques significantly improves the efficiency of the models, leading to substantial operational benefits.

The thesis is organized into seven chapters, each meticulously designed to address different facets of the research on predicting the mechanical properties of steel using machine learning.

In the second chapter, a comprehensive review of existing literature is presented, focusing on previous studies and advancements in machine learning applications in material science, particularly in predicting the mechanical properties of steel.

The Methodology chapter details the research methodology, including the machine learning models and feature selection methods used in the study. It begins with an overview of several machine learning models such as Linear Regression, Support Vector Regressor, Decision Tree Regressor, K-nearest neighbors, Random Forest, and eXtreme Gradient Boosting. Afterwards, it provides an overview of feature selection methods, such as Filter Methods, like Mutual Information Feature Selection and Anova F-Test Feature Selection, Wrapper Methods like Recursive Feature Elimination and Sequential Forward Selection, and Embedded Methods such as Least Absolute Shrinkage and Selection Operator.

The Numerical Experiments chapter includes the practical implementation of the research, including data collection, preprocessing, model training, feature selection, hyperparameter tuning, and the analysis of experimental outcomes. It begins with information about the data, followed by exploratory data analysis to understand its characteristics. The chapter describes the data preprocessing steps taken to prepare the data for modeling, the development of the machine learning models, the application of feature selection algorithms, and the hyperparameter tuning process to optimize the models. The outcomes of the experiments are analyzed and discussed, and the performance of the models is compared to that of predicting tensile strength and yield strength. It also examines the impact of feature selection on the models' performance, discusses the results for both tensile strength and yield strength predictions and concludes with an interpretation. The larger impact of the results is explored by comparing them with existing studies and discussing how they can be used in real-world situations. This looks at previous research to see how these findings fit into the larger field and how they can be used in practice.

The Conclusion chapter explains the main findings, why the research is important, and how it helps the field. It reviews the key conclusions and their importance for material science and machine learning. It also suggests possible improvements and areas that need more study. The chapter recommends how future work can build on these findings to move the field forward.

2. LITERATURE REVIEW

Recently, advances in artificial intelligence (AI), especially machine learning (ML) and deep learning (DL), have made people more interested in using data-driven methods in manufacturing and material science (Qian Xie et al. 2021). These methods have created new ways to analyze complex and nonlinear data in various research fields, such as material microstructures, inorganic nanomaterials, energy, and manufacturing 8Bhattacharyya et al. 2013). For a long time, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) have been used to predict material properties. They are good at handling complex problems. For example, they have been used to predict the properties of alloy steel based on its chemical makeup and hot-rolling process. While these methods work well with small datasets, their performance with large datasets still needs more study. Also, ANNs can have problems like overfitting and training difficulties when they are very deep. Additionally, ANNs can suffer from overfitting and convergence issues when they have many hidden layers (Krizhevsky, Sutskever, and Hinton 2012).

Besides DNN and CNN models, natural language processing (NLP) is also used to predict the mechanical properties of materials. This uses computational models that bring together chemical composition, manufacturing processes, and mechanical properties. NLP is utilized to transform qualitative manufacturing process data into a format suitable for neural networks to enhance the predictive accuracy of the models (A. P. O. Costa et al. 2024).

Zhang et al. proposed a CNN-based method to predict the mechanical properties of alloy steel using a new data preprocessing method that converts chemical composition and processing parameters into two-dimensional images for feature extraction (Z.-W. Xu, Liu, and Zhang 2019). The model is compared with traditional ANN and SVM methods on 60,000 industrial data points, showing promising results. This research offers significant guidance for optimizing steel compositions and production processes and developing new steel grades.

According to Xie et al., the core development phase for predicting the mechanical properties of hot-rolled materials involved creating a Deep Neural Network (DNN) model, which was trained using actual data from an advanced steel manufacturing plant. This model was designed to predict critical properties such as yield and tensile strength, incorporating 27 input features reflecting the steel's composition and production parameters. These parameters include plate dimensions, plate moving speed in the Accelerated Cooling Control (ACC) process, average cooling rate, start and finish cooling temperatures, and the flow ratios of upper and lower temperatures. Rigorous testing and optimization of the model's parameters ensured high prediction accuracy, significantly outperforming traditional methods.

Another study analyzed the dataset, including hot-rolled products such as S355, Q345, AH36, X80, 12MN, and Q550 (Li et al. 2020). For predictive performance, machine learning methods such as support vector machines, k-nearest neighbors, linear regression, and random forest were evaluated and compared with deep neural networks**.**

CNN predicts steel properties by turning production data into two-dimensional images. CNNs use fewer connections and shared weights, simplifying the model. They also use convolution and pooling to focus on local features, improving prediction accuracy. The tests show that the CNN model described is more accurate and robust than other models mentioned in the literature. This CNN model also helped us better understand the steel rolling process through sensitivity analysis (Z.-W. Xu, Liu, and Zhang 2019). The results matched the steel's known properties. For this reason, the CNN model helps predict the mechanical properties of hot-rolled steel products in real-world settings.

A recent study involved developing a method to predict the mechanical properties of TRIP (Transformation Induced Plasticity) aided steels using artificial neural networks (Bhattacharyya et al. 2013b). This approach integrates chemical composition and heat treatment data to predict the amount of retained austenite, which is crucial for achieving desired steel properties. ANN models, built on multilayer perceptron (MLP) architecture and trained using a scaled conjugate gradient backpropagation method, use inputs such as steel composition and heat treatment parameters, including inter-critical annealing and

isothermal bainitic transformation temperatures and times. The output is the volume fraction of retained austenite in the steel. Training and testing data are scaled and normalized to enhance processing efficiency and model accuracy.

Millner et al. applied AI regression models for predicting r-value, tensile strength, yield stress, and elongation at fracture of steel coils from chemical composition and process parameters. Various AI models are used in their study, such as Random Forest Regression, Support Vector Regression, Artificial Neural Networks, and Extreme Gradient Boost (Lugan et al., n.d.).

Despite the numerous studies utilizing neural networks for predicting the mechanical properties of materials, this thesis deliberately avoids using neural networks due to their inherent explainability issues. Neural network models are often considered "black boxes" making it difficult to understand how input parameters influence the output predictions. In a production environment where the results of this study will be applied, it is crucial to know which parameters affect specific mechanical properties. This transparency is essential for making informed decisions and optimizing the production process. Therefore, this study employs more explainable models like Linear Regression, SVR, Decision Trees, KNN, Random Forest, and XGBoost. These models provide more precise insights into the relationships between input features and predicted outcomes. That ensures the production team understands and trusts the model predictions.

3. METHODOLOGY

3.1. Overview of Machine Learning Models

3.1.1. Linear regression

Linear regression is an algorithm where the relationship between the inputs, which are the independent variables, and the outputs calculated based on these inputs is determined linearly (Su, Yan, and Tsai 2012), (Chou et al., n.d.). Linear regression establishes a relationship using the most suitable straight line between one or more independent variables (X) and the dependent variable (Y). The aim is to minimize the error value. Using the least squares method, a linear regression equation determines the β coefficients by minimizing the ε error. In linear regression, the dependent variable Y, the independent variable X, and the unknown parameters of this variable, β_1, and β_0, as well as the ε error term, are represented. β_0 indicates the point where the function intersects the y-axis, while β_1 represents the slope of the line. Linear regression analysis can be examined as simple and multiple linear regression. Simple linear regression is an algorithm where the relationship between the independent variable inputs and the calculated outputs based on these is determined linearly and can be defined as shown in equation (3.1):

$$Y=\beta_0+\beta_1 X_{1i}+\varepsilon_i \qquad i=1, 2,\ldots, n \qquad (3.1)$$

In multiple linear regression, there is a single dependent variable and several independent variables, and the linear relationship between them is expressed. In equation (3.2), X_1, X_2,…, X_n represents multiple variables, Y the dependent variable, and $\beta_0,\beta_1,\ldots,\beta_p$ the unknown parameters, along with ε representing the error. The multiple linear regression model is as given in equation (3.2) for p independent variables and n observations:

$$Y=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_p X_{pi}+\varepsilon_i \qquad i=1, 2,\ldots, n \qquad (3.2)$$

3.1.2 Support vector regressor

Support vector machines are techniques for both classification and regression that originate from statistical learning theory as introduced by Vapnik in 1995 (Pal, Singh, and Tiwari 2011). These methods employ the strategy of optimal class separation, choosing among countless potential linear classifiers to minimize the generalization error

or its upper limit based on the principle of reducing structural risk. For clearly separable classes, SVM selects a hyperplane that maintains the most significant possible margin between the classes, where the margin is the combined distances from the nearest class points to the hyperplane, as shown in Fig. 3.1. When classes overlap, SVM aims to optimize the margin while reducing misclassification errors. A preset positive constant manages the balance between margin width and misclassification. SVM's approach allows for adaptation to non-linear decision boundaries by mapping input variables into a higher-dimensional space, thereby transforming the problem into a linear classification in this new feature space.

The concept was further evolved with the introduction of support vector regression, which utilizes an e-insensitive loss function to find a function that deviates minimally from actual target vectors across all training data, striving for maximum flatness, as detailed by Smola in 1996. This extension includes the use of kernel functions to facilitate non-linear regression. The implementation of SVR involves fewer parameters set by the user. Beyond selecting a kernel, it necessitates adjusting kernel-specific settings, the regularization parameter C, and the error margin e in the sensitive zone, guiding the complexity of the model's predictions. A key strength of SVR is its optimization technique, which solves a linearly constrained quadratic programming problem, ensuring a unique, optimal, and globally applicable solution (Figure 3.1).

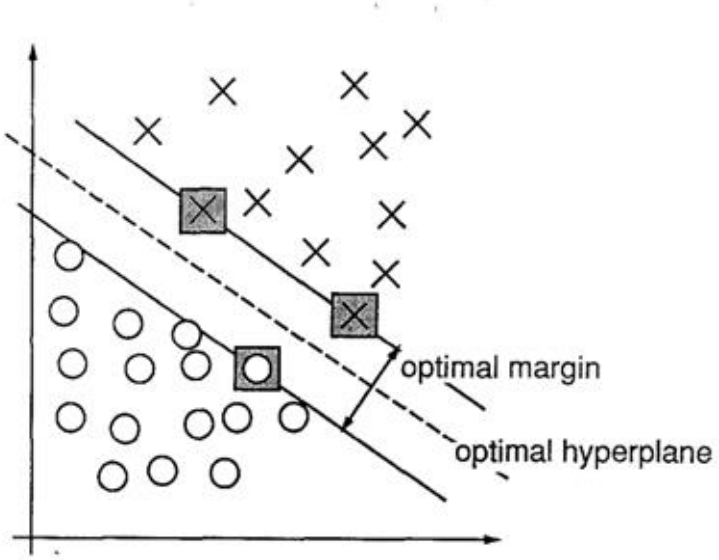

Figure 3.1 A scenario in a two-dimensional space where a problem is separable (Cortes and Vapnik 1995).

3.1.3. Decision tree regressor

Using decision trees in pattern recognition is highly effective for complex classification tasks. This nonparametric approach is also prevalent in machine learning due to its capability to automatically gather knowledge, which is essential for the creation of expert systems and knowledge-based systems (Sethi 1997).

Machine learning methods that utilize tree structures are often applied to engineering challenges to forecast results, whether for regression or classification purposes (M. Xu et al. 2005), (Pekel 2020). In the framework of decision trees, the data is recursively segmented based on rules that identify each node or branch. The objective of a decision tree is to create as homogeneous groups as possible at each node, effectively splitting the data from the root down to the leaves into the most advantageous subsets. Although primarily utilized for classification, this method is also adapted for regression tasks to predict numerical outcomes using a technique referred to as recursive partitioning. In such regression scenarios, the attributes being predicted are continuous variables. Decision tree regression algorithms employ a tree-based model to estimate the values of a dependent variable, where the decision-making elements are situated at the nodes, and the forecasted values are found at the leaves.

3.1.4. K-nearest neighbors

K-nearest neighbors are one of the oldest and easy-to-implement regression techniques (Tang and He 2015). In the KNN approach, all neighbors traditionally receive equal weight regardless of their similarity to the test instance. To address this limitation, it's beneficial to assign greater weights to neighbors that are more like the test instance. The weighting for each training instance can be calculated with a kernel function that is based on the distance—rather than similarity—from the test instance. The Euclidean distance metric is commonly used to calculate the distance between two instances (Nguyen, Morell, and De Baets 2016).

Hence, prior to utilizing the KNN method, it is crucial to scale the data. Subsequently, the KNN regressor determines the output value by averaging the data points that have

similar input features (Figure 3.2) (Morales-España, Mora-Flórez, and Carrillo-Caicedo 2010).

a) Data distribution and a new (Unknown) sample.

b) Prediction of the class of a new sample, considering the nearest neighbor

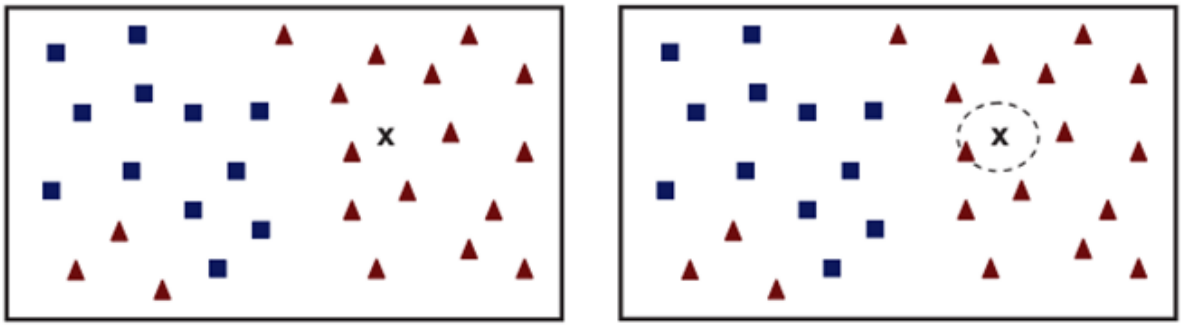

Figure 3.2 Determining the close points using the KNN method (Morales-España, Mora-Flórez, and Carrillo-Caicedo 2010)

3.1.5. Random forest

In machine learning, the Random Forest method is a detailed ensemble technique used mainly for classification and regression tasks. It involves the creation of multiple decision trees to improve the accuracy and consistency of predictions. Random Forest develops numerous decision trees and determines outputs during its training phase. Random forest identifies the most common class for classification tasks or computes the mean of the predictions for regression tasks (Breiman 2001).

This method is valued for its strong ability to avoid overfitting. The variety of trees in the Random Forest helps to reduce this issue. It handles large and complex datasets well and can find which features are important. This makes it a powerful tool for making predictions. (Hastie et al. 2009).

3.1.6. Extreme gradient boosting -XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a popular machine-learning algorithm for large datasets. It constructs a strong model through the integration of multiple decision trees. Essentially, XGBoost sequentially trains a.series of weak decision trees, learns from the errors of each and it enhances the model's accuracy (T. Chen and Guestrin 2016).

XGBoost starts by making initial predictions (usually zeros) for all instances. The algorithm adds new trees one at a time, each designed to correct the errors made by previous trees. This process involves sequentially building trees to minimize errors found in the dataset. At each step, the algorithm calculates the error rate (gradient) for each data point. These gradients help determine where the model is making errors and how to reduce these errors in subsequent steps. XGBoost controls the growth of trees to prevent overfitting. This involves pruning the trees at a certain depth or simplifying the tree structure. XGBoost includes regularization terms that penalize the complexity of the model. This ensures that the model generalizes well not only to the training data but also to unseen data. Predictions from all the trees are combined to produce the final model prediction (Chen and Guestrin 2016).

In machine learning, hyperparameter tuning is a critical step that involves adjusting the settings of an algorithm to optimize its performance. Below is a breakdown of the parameter grids defined for each model.

3.2. Overview of Feature Selection Methods

3.2.1. Filter methods

These methods use statistical techniques to score each feature in the data. Based on these scores, features are either kept or removed from the model. In the filter feature selection method, the choice of features is made without considering the model that will eventually use them. This is because filter models use the overall properties of the training data to choose features independently of any predictive model. They perform this selection as a preliminary step, not involving any machine learning algorithm directly (Sánchez-Maroño, Alonso-Betanzos, and Tombilla-Sanromán 2007).

Filter methods are quicker than wrapper methods and often lead to models that generalize better because they do not depend on the model used for making predictions. However, these methods might choose too many features, sometimes even all of them, which requires setting a limit to pick a smaller, more relevant set of features (Sánchez-Maroño,

Alonso-Betanzos, and Tombilla-Sanromán 2007). In this study, the following two different feature selection methods have been applied as filtering methods.

3.2.1.1. Mutual Information Feature Selection

Regarding the Mutual information-based feature selection algorithm, Battiti introduced a feature selection algorithm in 1994, which aims to pick features that strongly relate to the output while avoiding redundancy (Amiri et al. 2011). An operating classifier, like a multilayer perceptron using the backpropagation algorithm, is seen as a system that reduces initial uncertainty by processing the information in the input vector (Rumelhart, Hinton, and Williams 1986). Ideally, this process would remove all uncertainty, ensuring the class is clearly identified. However, in real-world scenarios, some uncertainty usually remains due to two main reasons: either there isn't enough input information, or the system isn't operating optimally.

In cases where the system is suboptimal, despite having enough information, it might waste some due to not being fully trained or because of certain approximations or failures. This can often be improved by using more training examples, extending the training period, or trying different algorithms. On the other hand, if the problem is a lack of sufficient input information, it’s crucial to identify this early in the development process. The solution here would typically involve adding more or better features to the dataset (Battiti 1994).

3.2.1.2. Anova F-Test Feature Selection

The ANOVA F-Test in feature selection is a statistical approach used to identify the most important features that show a strong correlation with the output variable. ANOVA stands for Analysis of Variance, and it's commonly used to assess whether the differences in group means are statistically significant. This method helps compare the means from different groups to see if there are any meaningful differences among them.

The ANOVA F-Test feature selection method, as detailed in the provided article, is a statistical approach used to identify the most significant features affecting the outcome of a model. This method uses the one-way ANOVA (Analysis of Variance) F-Test to

check which features are strongly connected to the dependent variable. The ANOVA F-Test compares the averages of different groups to see if the differences are important. In feature selection, this helps reduce the number of features before classification. This makes the model simpler and improves its performance. The ANOVA F-Test is useful for finding which features help the most in separating different classes. It is helpful in tasks like spam email detection, where it can improve a classifier like SVM by finding and keeping the most important features (Elssied, Ibrahim, and Osman 2014).

3.2.2. Wrapper methods

Wrapper methods in feature selection assess the quality of different feature subsets by using the performance of a specific modeling algorithm, which is considered a black box. For classification, for instance, a wrapper might measure the effectiveness of subsets using classifiers like Naïve Bayes or SVM (Bradley and Mangasarian 1998), (Maldonado, Weber, and Famili 2014). In clustering tasks, it evaluates subsets based on how well they work with algorithms such as K-means. Each subset is repeatedly evaluated, and its generation depends on the chosen search strategy, similar to how filters operate.

However, wrappers tend to be slower than filters because they rely on the computational demands of the modeling algorithm used. The selected feature subsets are also tailored to the specific modeling algorithm, potentially leading to biased results unless cross-validation is employed. To ensure a reliable estimate of the generalization error, it's advisable to use an independent validation sample and possibly a different modeling algorithm after the best subset is identified. Despite these challenges, wrappers often lead to better-performing subsets than filters since they assess subsets using actual modeling algorithms. While various combinations of search strategies and modeling algorithms can be implemented as wrappers, they are most effective with greedy search strategies and fast algorithms like Naïve Bayes, linear SVM, and Extreme Learning Machines (Jović, Brkić, and Bogunović 2015).

3.2.2.1. Recursive Feature Elimination

Recursive Feature Elimination used for feature selection in machine learning. The traditional RFE method works by iteratively removing the least significant features based

on their impact on the model's performance, starting with the smallest weights and moving on to larger weights (Hastie et al. 2009). This is known as recursive feature elimination, usually with support from models like Support Vector Machines. This methodology is especially useful in scenarios with small sample sizes but high dimensionality, where traditional methods might overlook the combined potential of features that appear weak when isolated (X. Chen and Jeong 2007).

3.2.2.2. Sequential Forward Selection

Wrapper-type approaches use the classification performance of the classifiers as a numerical evaluation. Sequential Forward Selection is a wrapper type of feature selection method.

SFS begins with an empty set of features and progressively adds the most accurate features, as determined by their impact on classification performance until all features have been evaluated (Chandrashekar and Sahin 2014). This method enhances model performance by including only the most relevant features, which reduces computational complexity and potentially increases the detection rate of the system (Lee, Park, and Lee 2017).

3.2.3. Embedded methods

Embedded methods integrate feature selection directly into the model training process, effectively making it part of the algorithm's core or extended functionality. These methods are particularly common in decision tree algorithms such as CART, C4.5, and Random Forest (Sandri and Zuccolotto 2006). However, they are also used in other types of models like multinomial logistic regression (Cawley, Talbot, and Girolami 2006).

In embedded methods, feature selection is achieved by adjusting the model to not only minimize prediction errors but also to simplify the model. This is done by penalizing or reducing the coefficients of less important features to zero. Techniques like Lasso (Ma and Huang 2008) and Elastic Net (Zou and Hastie 2005) are examples of this approach; they apply regularization, which adds a penalty to the model's loss function based on feature coefficients. This encourages the model to consider fewer features, which can be

particularly effective with linear classifiers such as SVM. These methods help enhance model performance by keeping only the most relevant features, thus reducing overfitting and improving model generalizability (Jović, Brkić, and Bogunović 2015). Additionally, Random Forest and XGBoost are categorized in the embedded methods (Chemmakha, Habibi, and Lazaar 2022)

3.2.3.1. Least Absolute Shrinkage and Selection Operator – Lasso

LASSO, which stands for Least Absolute Shrinkage and Selection Operator, was developed by Robert Tibshirani in 1996. This method performs two main tasks by limiting the sum of the absolute values of model parameters: regularization and feature selection. The sum must be less than a specified upper limit, applying a penalty process that shrinks some coefficients to zero. During this process, variables that still have a non-zero coefficient after shrinking are chosen to be part of the model (Fonti and Belitser 2017). Practically, the tuning parameter λ is crucial; when λ is sufficiently large, coefficients are reduced to zero, facilitating dimension reduction. Conversely, when $\lambda=0$, the method reverts to ordinary least squares regression (OLS). LASSO offers several advantages, particularly improving prediction accuracy by reducing variance without significantly increasing bias, which is especially valuable when the number of observations is small and the number of features is large. It also enhances model interpretability by eliminating irrelevant variables not associated with the response variable, thereby reducing model overfitting. This paper focuses on the feature selection task, which is a key interest area in this study (Fonti and Belitser 2017).

3.2.3.2. Random Forest

Random Forest is a flexible and user-friendly supervised learning algorithm that effectively handles classification tasks without the need for many hyper-parameters. In Random Forest, it's essential to construct a minimum number of trees to ensure all data is classified, and this number largely depends on the specific dataset. The 'breaker attributes'—or key features—significantly influence how many trees are needed. Generally, as the number of trees increases, so does the model's accuracy. However, there is a point of maximum accuracy beyond which adding more trees does not improve results. Accuracy also depends on the number of breaker attributes used; using a number

equal to the total attributes available typically results in lower accuracy (Huljanah et al. 2019).

3.2.3.3 XGBoost

The XGBoost feature selection model described in this article is used during machine learning models' training process to improve efficiency and performance. By measuring the gain of each feature used in the model, it determines the most impactful features for model decisions. This gain is calculated from the improvement in accuracy brought by a feature to the splits it is used in. A feature's total gain across all trees in the model is divided by the number of times it was used for splitting, thus assigning it an importance score. Features with higher scores are deemed more significant, influencing the construction and training of the model. This systematic approach helps in selecting the most useful features, enhancing the model's accuracy and efficiency (Jiang et al. 2023).

3.3. The Proposed study

In this subsection, the proposed methodology employed for predicting the mechanical properties of steel using machine learning is detailed. Initially, **Data Preprocessing** is conducted to clean and prepare the raw data for analysis. Subsequently, **Model Development** is undertaken to construct various machine learning models aimed at predicting the target properties. Following this, **Hyperparameter Tuning** is meticulously performed to enhance model efficacy. **Feature Selection** techniques are then applied to identify the most influential features. A rigorous **Performance Comparison** is conducted to evaluate the models, culminating in **Model Selection,** where the optimal model is chosen. The general flow of this methodology is depicted in the accompanying Figure 3.3, providing a visual representation of the sequential steps and their interconnections.

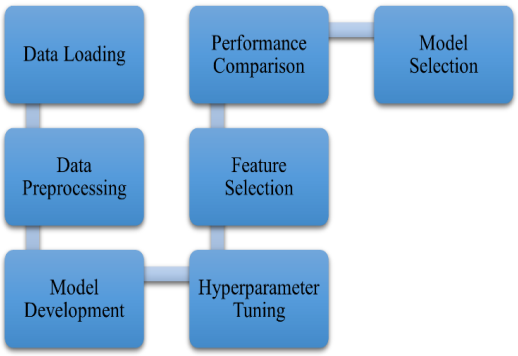

Figure 3.3 Flowchart of the Machine Learning Model Development Process

3.4. Information about the data

In this study, the mechanical properties of coils, specifically yield strength and tensile strength measured with a tensile test device, have been utilized to create a comprehensive database spanning the years 2018-2024. The testing system is illustrated in Figure 3.4.

Figure 3.4 Modern closed-loop servo-hydraulic testing system (It was taken in Borcelik Testing Laboratories)

A tension test consists of slowly pulling a sample material with an axial until it breaks. This is a destructive test method.

In flat steel manufacturing, the test specimen used has a rectangular cross-section; its ends are usually enlarged to provide extra area for gripping and to avoid having the

sample break where it is being gripped. The usual manner of conducting the test is to deform the specimen at a constant speed. An axial force that must be applied to achieve this displacement rate varies as the test proceeds. This force P may be divided by the cross-section area Ai to obtain the stress in the specimen at any time during the test (E. Dowling Norman, Siva Prasad Katakam, and R. Narayansamy 2013)

Within this study, a total of 24 input parameters, comprising chemical composition-related parameters and process parameters, have been employed for use in machine learning algorithms. The following table 3.1 provides a summary of the input features included in the dataset.

Table 3.1 Input Features of Mechanical Property Machine Learning Models for Cold Rolled Steel Production Process

Input Features	
Annealing Temperature	Nb (%)
Soaking Time	P (%)
Annealing Weight	S (%)
Al (%)	Si (%)
B (%)	Ti (%)
C (%)	V (%)
Ca (%)	Ceq
Cr (%)	Reduction rate
Cu (%)	Thickness
Mn (%)	Width
Mo (%)	Skinpass Elongation
N (%)	Skinpass Rollforce

In this study, the impact of chemical properties as input parameters on the mechanical characteristics of steel is examined. Specifically, the roles of Chromium, Vanadium, Manganese, Titanium, Niobium, Silicon, Carbon, Copper, Sulfur, Phosphorus, Molybdenum, Nitrogen, Boron, Calcium, and Aluminum in determining the final mechanical properties of cold-rolled flat steel are investigated. Through rigorous analysis

and modeling, a comprehensive understanding is sought of how these chemical elements individually and collectively influence the steel's tensile strength and yield strength (A. Costa et al., n.d.).

Carbon Equivalent is one of the input parameters that will be used within the machine learning algorithms in this study. Carbon Equivalent (CE) in equation 3.3 is a parameter used to assess the weldability of steel, indicating how the combination of various alloying elements in the steel affects its hardness and susceptibility to cracking during welding. Although CE primarily concerns weldability, it indirectly relates to the mechanical properties of steel because the weldability of a material can influences its overall strength, toughness, and ductility post-welding .

$$\text{CEV} = \text{C} + \frac{Mn}{6} + \frac{(Cr+Mo+V)}{5} + \frac{(Ni+Cu)}{15} \quad (3.3)$$

- C is the carbon content in percentage,
- Mn is the manganese content in percentage,
- Cr is the chromium content in percentage,
- Mo is the molybdenum content in percentage,
- V is the vanadium content in percentage,
- Ni is the nickel content in percentage,
- Cu is the copper content in percentage.

Thickness and width are the dimensions of the coils.

Skin Pass Elongation: The "skin pass" line is the final step in the cold-rolled flat steel manufacturing process. This line produces lightly rolling steel to improve its surface finish and mechanical properties. The elongation value related to the skin pass line, referred to as "skin pass elongation" or "temper rolling elongation," refers to the slight increase in the length of the steel sheet due to this process. This deformation, about less than 4%, improves the steel's surface texture and flatness and imparts the work hardening. The skin pass elongation helps to the steel's final mechanical properties by increasing its

yield strength slightly and improving its formability, making it more suitable for subsequent forming operations (Lugan et al., n.d.; Technology and 2015, n.d.).

Skin Pass Line Force: The "roll force" in the skin pass line refers to the mechanical force the rollers apply onto the steel sheet during this process. This force is crucial for achieving the desired amount of deformation, influencing the final surface texture and flatness, and ensuring the steel has the required mechanical characteristics for its intended application (Technology and 2015, n.d.).

Reversible Cold Line Reduction Rate: The rolling reduction ratio is obtained by subtracting the target rolling thickness from the raw materials (hot roll) thickness and dividing the result by the raw material's thickness. Through the rolling process, the hot roll coil or product is mechanically destroyed, resulting in a semi-finished product known as FH (full hard). The annealing process is then required for further processing (Ahmad et al. 2014).

Annealing Temperature: Annealing is a heat treatment process used to reduce hardness, increase ductility, and help eliminate internal stresses within the steel. For flat steel, as well as many other steel types, the annealing temperature can typically range from around 600°C to 730°C, although the exact temperatures can vary based on the steel's composition and the specific goals of the annealing process (such as stress relief, recrystallization, or full anneal).

Annealing Soaking Time: During batch annealing, the term "soaking" refers to the total annealing time at the maximum annealing temperature in the batch annealing furnace.

Batch Annealing Furnace Base Weight: Batch annealing base weight represents the total annealing tonnages of the coils loaded in the batch annealing furnace.

4. NUMERICAL EXPERIMENTS

In this research, several Python libraries were utilized for feature selection algorithms, encompassing a range of methods to ensure comprehensive analysis and selection of the most relevant features. The scikit-learn library was employed extensively. The RandomForestRegressor was used for its ensemble methods, which are crucial for determining feature importance. Additionally, scikit-learn's f_classif and mutual_info_regression functions were applied to perform feature selection based on statistical tests and mutual information criteria, aiding in identifying significant features. The Lasso method from scikit-learn was also utilized, which helps select features by penalizing the absolute size of the coefficients, effectively reducing less essential features to zero and thus selecting a subset of the most relevant features. The mlxtend library was explicitly used for its SequentialFeatureSelector (SFS) class. This class enables sequential forward and backward selection of features, providing a systematic approach to feature selection that is both comprehensive and efficient. Lastly, the XGBRegressor from the xgboost library was used. It is known for its high performance in regression tasks through gradient boosting techniques. This method provides robust feature selection capabilities by evaluating the importance of each feature in the predictive model.

The computational environment for this study consisted of a system with an 11th Gen Intel(R) Core(TM) i7-1165G7 processor, operating at 2.80 GHz and equipped with 16 GB of RAM (15.7 GB usable). The system is a 64-bit operating system on an x64-based processor, ensuring sufficient computational power and memory to handle the data processing and model training tasks required for this research.

4.1. Exploratory data analysis

Let X be the dataset containing 5000 samples with 24 features. This dataset also has two outputs, namely yield and tensile strengths. All features in the dataset are of float type. The input features include various process parameters and the ratios of different chemical elements. This detailed examination is conducted to uncover patterns, trends, and relationships within the data, providing insights into the factors influencing the mechanical properties of the materials.

Initially, the general correlation of the input features was examined using a correlation heatmap (Figure 4.1). The output features were found to be correlated with each other, with a correlation of 0.9 between tensile strength and yield strength. Additionally, the carbon equivalent feature was observed to be correlated with both yield and tensile strength. It was also noted that the chemical features exhibited weak correlations with each other.

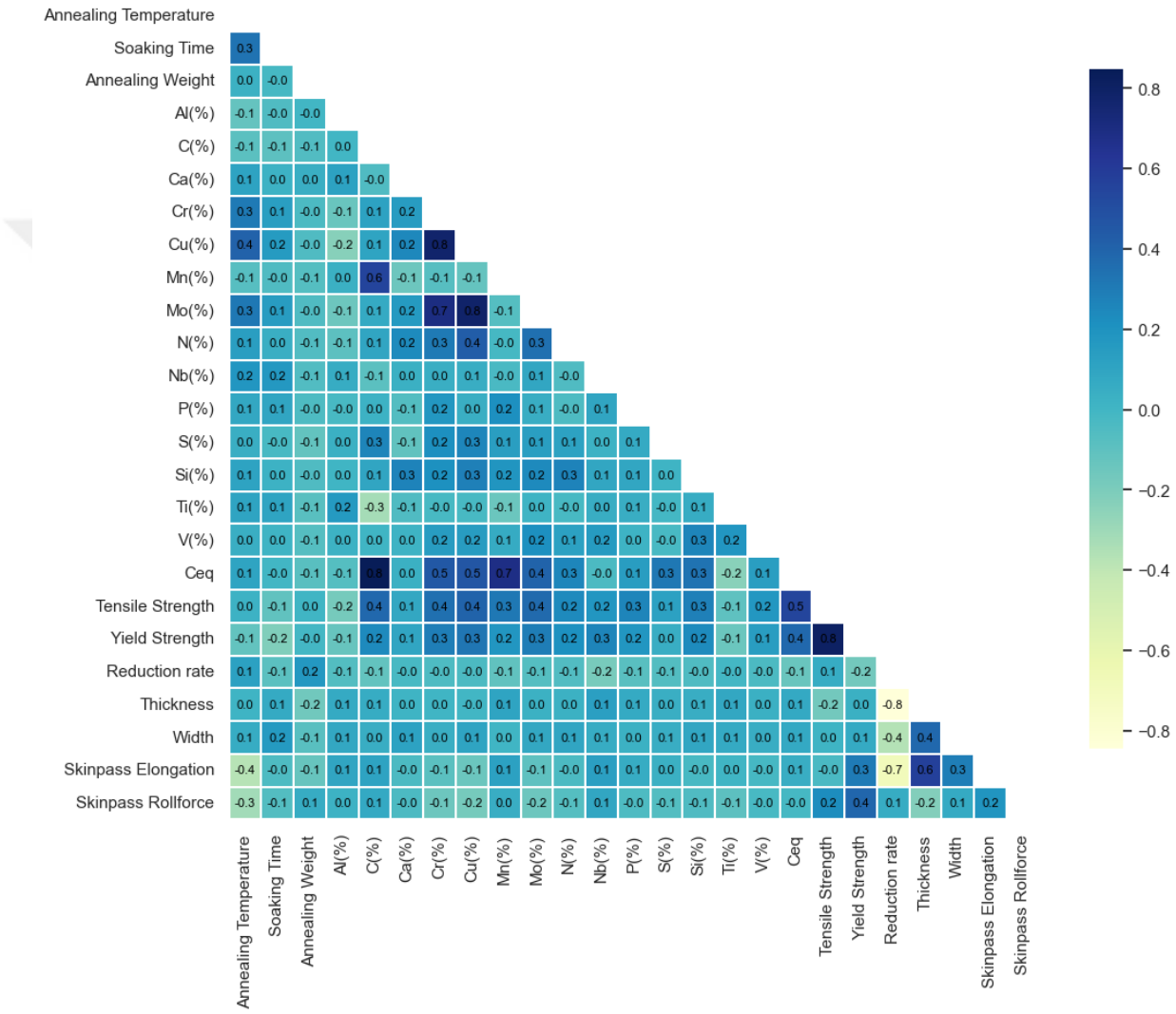

Figure 4.1 Correlation Matrix

The distribution plots of yield strength and tensile strength (Figure 4.2), as well as a scatter plot, were generated. In the scatter plot (Figure 4.3), the high correlation between the two features can be clearly seen.

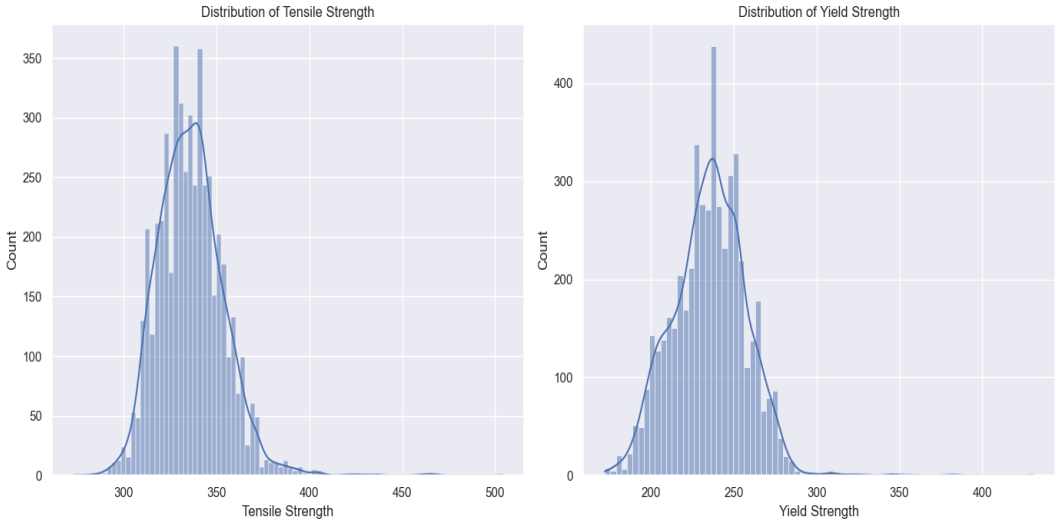

Figure 4.2 Histogram of Yield Strength and Tensile Strength

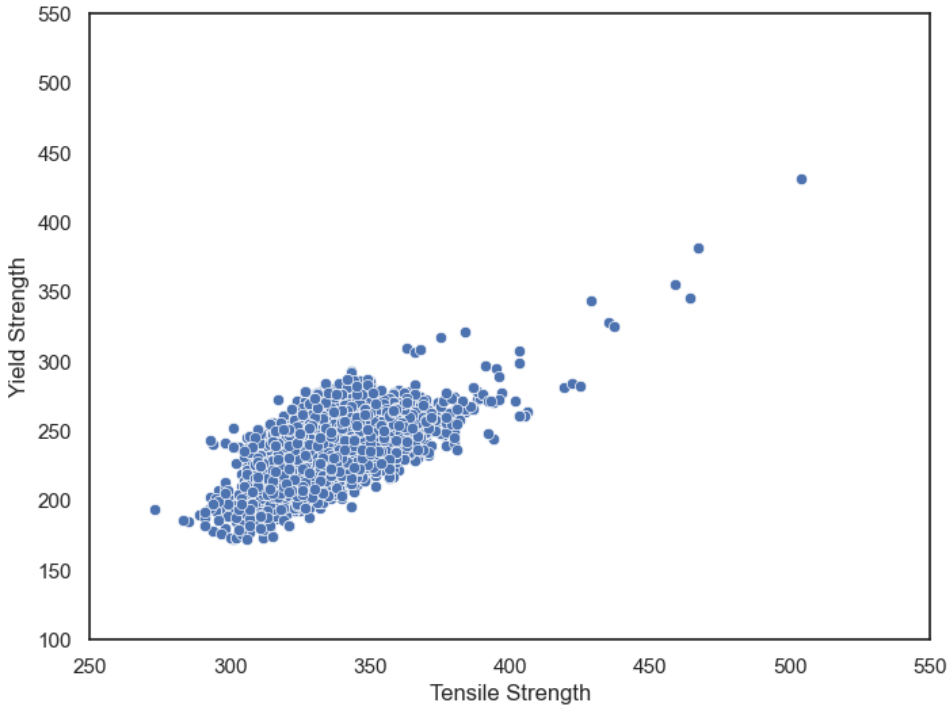

Figure 4.3 Scatter Plot of Yield Strength vs. Tensile Strength in Cold Rolled Coils

The distributions of Mn%, C%, Ti%, and P% values were also plotted (Figure 4.4). Although these features exhibit small-range changes, they could have a significant effect on the output features.

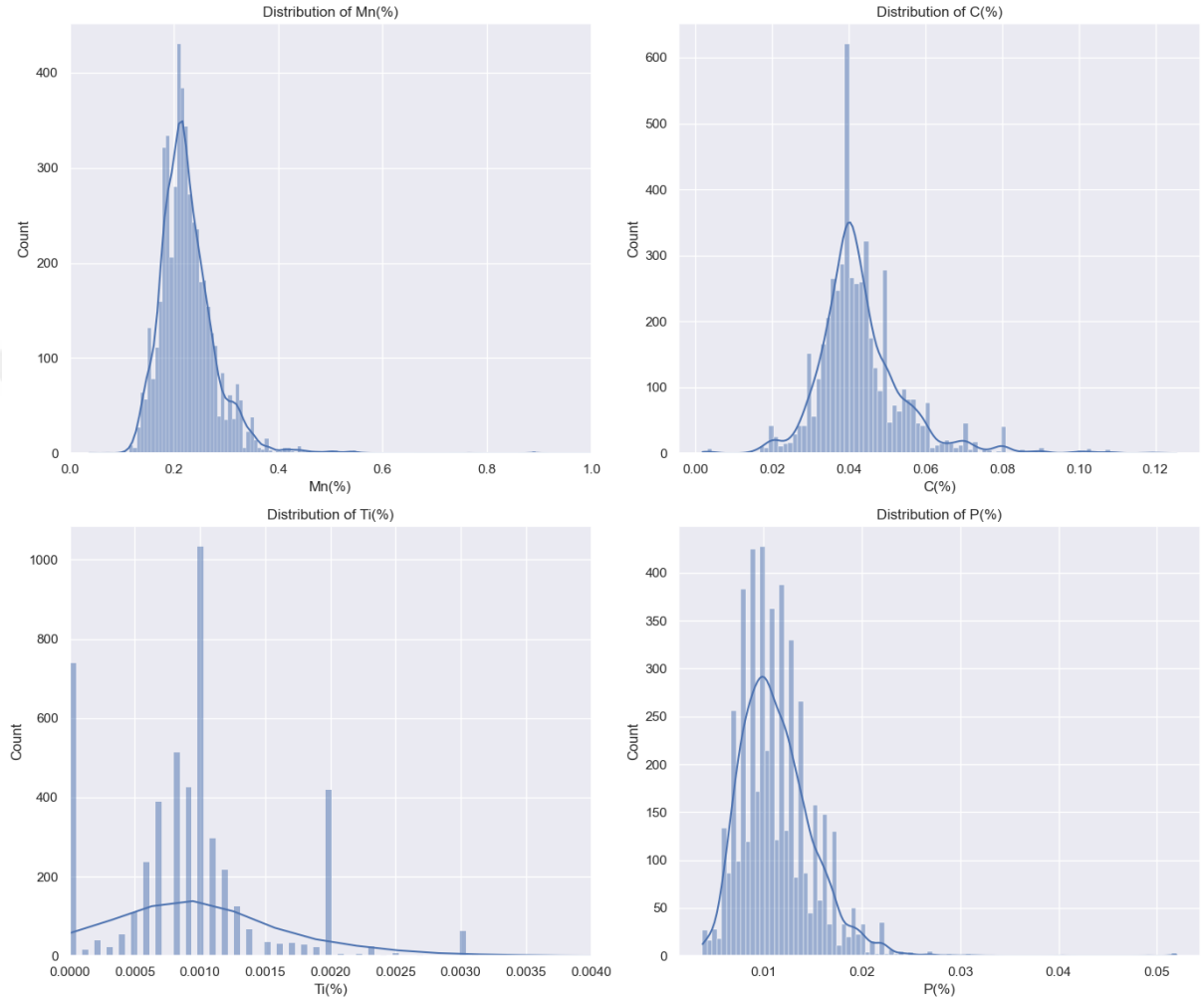

Figure 4.4 Distribution of Chemical Composition in Cold Rolled Coils

This dataset focuses on three different steel quality groups: DC01, DC03, and DC04. The distribution of these quality groups within the dataset was visualized using a bar plot (Figure 4.5).

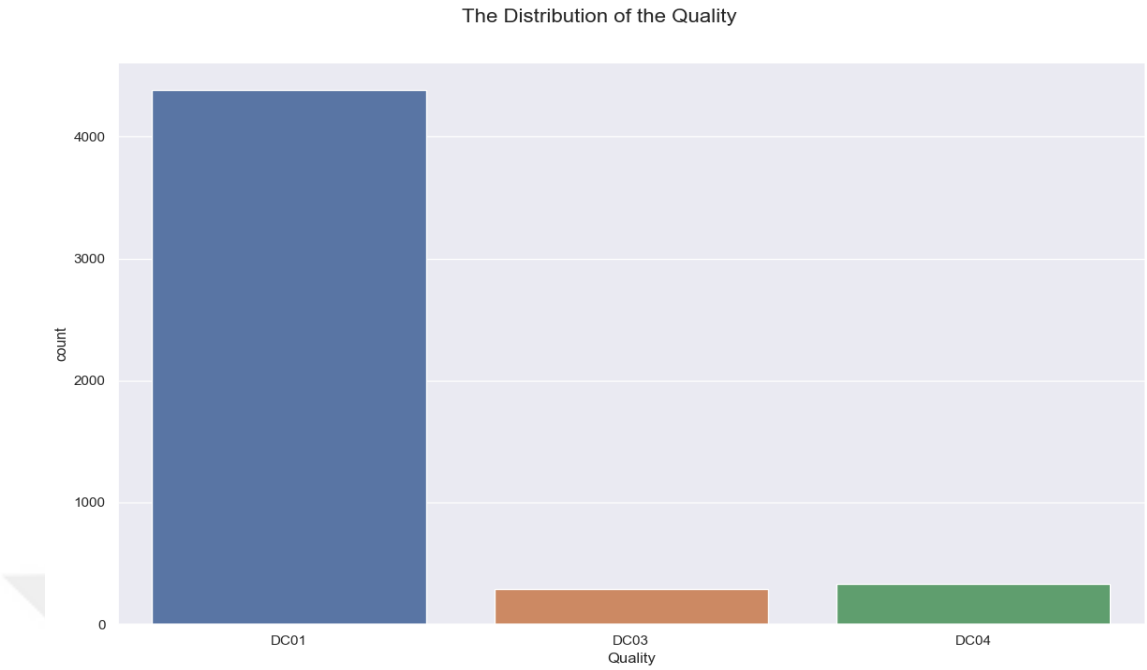

Figure 4.5 Distribution of Quality in Data Frame

In this dataset, data are derived from two distinct supplier groups: Japanese suppliers and European suppliers. The distribution of yield strength and tensile strength values for these supplier groups is illustrated in the following distribution plots (Figure 4.6 and Figure 4.7). These visualizations provide insight into the comparative mechanical properties of steel from different geographical sources, facilitating a deeper understanding of the material characteristics influenced by supplier origin.

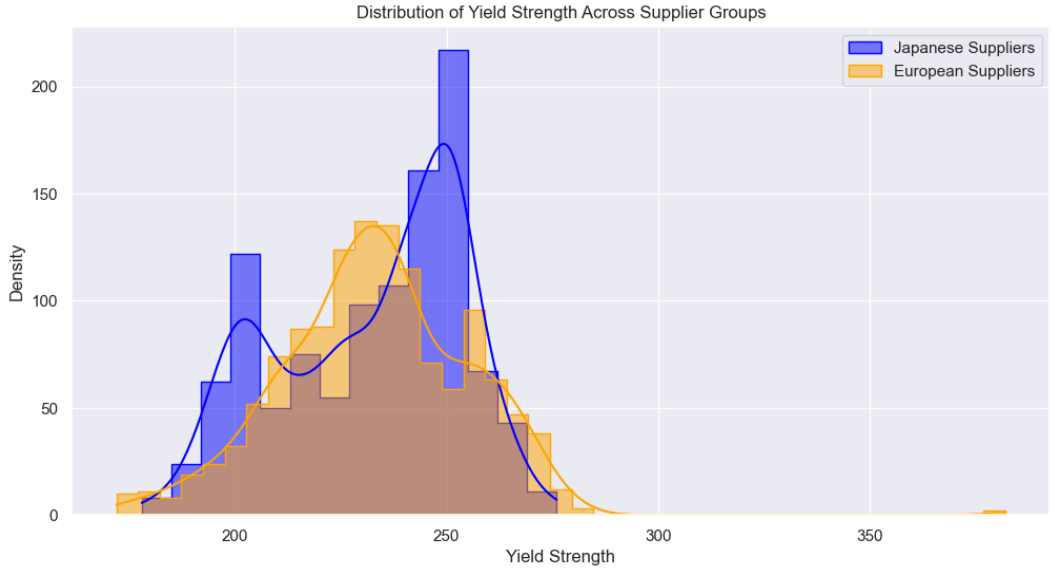

Figure 4.6 Histogram of Yield Strength Across Supplier Groups

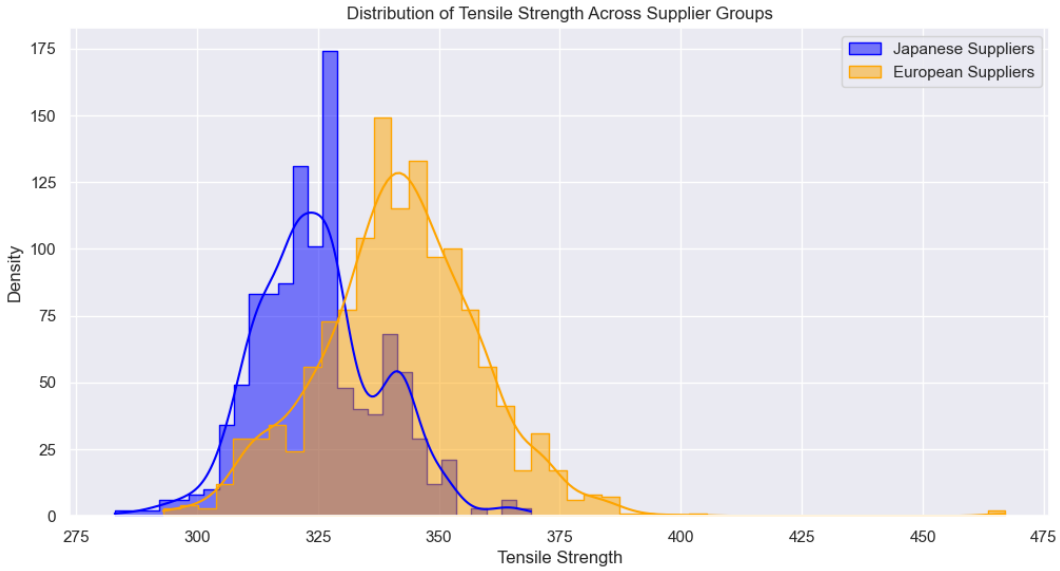

Figure 4.7 Distribution of Tensile Strength Across Supplier Groups

4.2. Data preprocessing

Data cleansing also referred to as data cleaning, was performed to detect and correct inaccuracies and inconsistencies in the dataset, thereby enhancing its quality. This process encompassed addressing a range of data issues, such as duplicates, missing values, and inconsistencies in data format, among others (Ridzuan and Zainon 2019). For instance, data duplication or absence could skew statistics, leading to unreliable or

misleading conclusions. The complexity of data cleaning stemmed from the diversity of data inconsistencies and the volume of data, making it a challenging aspect of data management.

Following the data cleansing process, the standard scaling preprocessing method was applied to normalize the attributes of the dataset (Hackeling 2017). Standard scaling is utilized to normalize the attributes by eliminating the average and adjusting each attribute to have a unit variance. This normalization is crucial as it enhances the performance of several machine learning models that depend on uniformity in feature scale. Specifically, the data were standardized by setting the mean of each attribute to zero and the variance to one. By subtracting the mean and dividing by the standard deviation for each feature, the Standard Scaler ensured an even contribution of each feature to the model's analysis. This technique is widely adopted across numerous machine-learning algorithms due to its effectiveness in feature scaling. The equation 4.1 is used for standard scaling. Here, equation 4.1 represents the original data point, μ is the mean of the data, and σ is the standard deviation of the data.

$$z = \frac{x-\mu}{\sigma} \tag{4.1}$$

4.3. Model development and feature selection

In this study, two distinct target variables within a data frame were examined: Yield Strength (Mpa) and Tensile Strength (Mpa). Yield Strength (Mpa) represents the material's yield strength, while Tensile Strength (Mpa) denotes the tensile strength of the material. Following the identification of these target variables, the data were partitioned into training and testing subsets with an 80/20 split (x_{tr} , x_{test} , y_{tr} , y_{test}). x is the one of the features of X and y is the target-dependent variable. This division was critical for assessing the models' generalization capabilities.

x_{tr} *= The training data for the features (independent variables). It is the portion of the dataset used to train the machine learning model*

x_{test} *= This is the test data for the features. It is the portion of the dataset used to evaluate the performance of the trained machine-learning model.*

y_{tr} = *This is the training data for the target variable (dependent variable). It corresponds to the output values that the model needs to predict during the training phase.*

y_{test} = *The test data for the target variable. It corresponds to the actual output values used to compare against the model's predictions during the evaluation phase.*

Six distinct regression models were defined: Linear Regression, Support Vector Regressor, Decision Tree, K-Nearest Neighbors, Random Forest, and eXtreme Gradient Boosting. Parameter options for each model were then established, including the hyperparameters used for optimizing the models. For these models, grids of parameters were defined for tuning, specifying different configurations to test during the model tuning phase, such as the number of estimators in a forest or the depth of a tree. hyperparameter tunning with a cross-validation strategy was utilized to find the best model parameters based on the negative mean squared error. The performance of the models was quantified using the R^2 metric.

4.4. Hyperparameter tuning

Support Vector Regression: The parameter C trades of the correct classification of training examples against the maximization of the decision function's margin, with values of 50, 100, and 150 to balance between bias and variance. The parameter gamma defines the influence of a single training example, with values of 05, 0.1, and 0.15, to adjust the decision boundary flexibility. The kernel type 'rbf' (radial basis function) is used for non-linear data.

Decision Tree Regressor: The maximum depth of the tree, with values of 30, 40, and 50, limits the number of nodes to balance model complexity and generalization. The minimum number of samples required to split an internal node is set to 2, encouraging deeper trees. The minimum number of samples required to be at a leaf node is set to 1, allowing leaves to hold as few as one sample.

K-Nearest Neighbors Regressor: The number of neighbors that vote on the classification, tested with values of 2, 3, and 4, helps find the optimal neighborhood size. The weight function (weights) 'distance' is used, giving greater influence on closer neighbors. The Euclidean metrics are utilized points.

Random Forest Regressor: The number of trees in the forest tested with values of 25, 50, and 75, explores the effect of ensemble size. The number of features to consider when looking for the best split is set to 'sqrt'. The maximum depth, minimum number of samples required to split an internal node and the minimum number of samples required to be at a leaf node are similar to those in the Decision Tree model.

XGBoost Regressor: The number of gradient-boosted trees is varied with values of 150, 200, and 250. The learning rate values of 0.5, 0.1, and 0.15 help prevent overfitting by making the boosting process more conservative. The tree depth values of 7, 9, and 11 allow the model to fit complex patterns. The fraction of features to use per tree values of 0.8 and 1, and the fraction of samples used for each tree values of 0.7, 0.8, and 0.9, ensure variability and robustness in training.

These grids were utilized in a randomized search to experimentally determine the best parameters for each model, aiming to improve prediction accuracy while managing overfitting. A function was defined to fit each model on the training data, predict the test data, and compute evaluation metrics. Another function employed a randomized search strategy with cross-validation (using K-Fold) to find the optimal model parameters based on the negative mean squared error. This process aimed to optimize the settings of machine learning models by identifying the best hyperparameters, leveraging the capabilities of the randomized search method from Scikit-learn.

Combining K-Fold cross-validation with a randomized search method enhanced the robustness of the hyperparameter tuning process. Each set of hyperparameters selected by the randomized search was evaluated across multiple folds, optimizing data usage and ensuring robust hyperparameter tuning. K-Fold cross-validation divided the dataset into Kequal parts, using each part once as a test set and the remaining parts as the training set. This method assessed the model's performance across different data subsets, ensuring robust model evaluation.

The randomized search method selected hyperparameter combinations from a defined pool, testing each combination on the training and test sets generated by K-Fold cross-validation. This process ensured that the model fitted well to specific data subsets and

across the entire dataset, enhancing the model's generalization ability and reducing the risk of overfitting (Géron 2022).

4.5. Feature selection

To improve the R^2 score performance of the models, seven different feature selection methods including filtering, wrapper, and embedded feature selection approaches—were applied to 24 distinct features.

Feature importance refers to techniques that assign a score to input features based on how effective they are at predicting a target variable (König et al. 2021). It plays an essential role by offering insight into the data and model. Therefore, it is possible to increase a model's performance and efficiency by using the feature selection process.

Feature selection is crucial in classification because including features that are irrelevant or redundant can often slow down the process and reduce the accuracy of the predictions made by a classification algorithm. A feature might seem irrelevant by itself if it doesn't correlate well with the class label. However, when used together with other features, it can become very important. If these features are mistakenly removed, it could lead to losing valuable information, which might result in a worse performance of the classification model (El Akadi, El Ouardighi, and Aboutajdine 2008). To enhance the accuracy of the model, seven different feature selection methods are applied. Each method serves to identify the most relevant features from the dataset, reducing dimensionality and improving model performance. In Table 4.1, a brief explanation of the used feature selection technique is mentioned:

Table 4.1 Feature Selection Methods applied to Machine Learning Models to improve accuracy.

Feature Selection Methods	
Filter Methods	Mutual Information
	Anova F-Test
Wrapper Methods	Recursive Feature Elimination
	Sequential Forward Selection
Embedded Methods	LASSO
	Random Forest
	XGBoost

4.6. Results and analysis

In this research, it is focused on the prediction of yield and tensile strengths using various feature selection methods across different datasets. Comparing Table 4.2 with Table 4.3, a significant improvement in the performance of predictions related to yield and tensile strength has been observed in both data and geography-based datasets.

Table 4.2 Comparison of R^2 Values for Mechanical Property Modeling Using Various Machine Learning Methods across Diverse Geographic Data Sets

	YIELD STRENGTH			TENSILE STRENGTH		
	ALL	EU	JAPON	ALL	EU	JAPON
Linear Regression	0,713	0,778	0,630	0,602	0,637	0,600
SVR (Tuned)	0,902	0,928	0,898	0,858	0,860	0,859
Decision Tree (Tuned)	0,878	0,889	0,836	0,799	0,766	0,779
KNN (Tuned)	0,918	0,927	0,919	0,880	0,884	0,896
Random Forest (Tuned)	0,931	0,941	0,910	0,898	0,878	0,885
XGBoost (Tuned)	0,938	0,950	0,918	0,908	0,879	0,895
Target	**0,950**	**0,950**	**0,950**	**0,950**	**0,950**	**0,950**

Table 4.3 Comparison of Feature Selection Methods of R^2 Values for Mechanical Properties

		Yield Strength			Tensile Strength		
		All	EU	Japan	All	EU	Japan
Filter Methods	Mutual Information	94.1 (XG) (23^{rd})	95.3 (XG) (15^{th})	92.5 (XG) (21^{st})	91.2 (XG) (23^{rd})	89.5 (KNN) (22^{nd})	90.1 (XG) (23^{rd})
	ANOVA F-Test	94.1 (XG) (23^{rd})	95.2 (XG) (24^{th})	92.1 (XG) (22^{nd})	91.2 (XG) (23^{rd})	88.4 (KNN) (24^{th})	90.3 (XG) (22^{nd})
Wrapper Methods	RFE	94 (XG) (22^{nd})	95.2 (XG) (24^{th})	92.1 (XG) (21^{st})	91.2 (XG) (22^{nd})	88.6 (KNN) (21^{st})	89.7 (XG) (21^{st})
	SFS	94 (XG) (18^{th})	95.3 (XG) (24^{th})	94.5 (XG) (23rd)	91.1 (XG) (24^{th})	88.4 (KNN) (24^{th})	91.6 (XG) (22^{nd})
Embedded Methods	LASSO	93.8 (XG) (24^{th})	95.4 (XG) (24^{th})	92.4 (XG) (22^{nd})	90.8 (XG) (23^{rd})	88.8 (KNN) (19^{th})	89.9 (XG) (21^{st})
	Random Forest	93.8 (XG) (20^{th})	95.3 (XG) (24^{th})	92.3 (XG) (22^{nd})	90.9 (XG) (23^{rd})	88.4 (KNN) (24^{th})	89.9 (XG) (24^{th})
	XGBoost	93.8 (XG) (20^{th})	95.3 (XG) (19^{th})	92.4 (XG) (22^{nd})	90.8 (XG) (21^{st})	88.4 (KNN) (24^{th})	90.2 (XG) (24^{th})

Tables 4.2 and 4.3 summarize the R-squared (R^2) values for yield and tensile strength across six different models using datasets from all sources, EU and Japan data. The results indicate significant variations in model performance based on the geographic origin of the data and the choice of models.

For yield strength, the models perform best when utilizing all available data, with the XGBoost model achieving the highest R^2 value of 93.8%. This suggests that a more comprehensive dataset provides a richer basis for the model to accurately understand and predict material behaviors. The EU data closely follows, with the XGBoost again showing superior performance with an R^2 of 95%. This high value reflects the model's robustness and the quality of the EU dataset in capturing the essential dynamics of yield strength. On the other hand, the Japan data, while slightly lower, still shows commendable results, especially with the KNN model, which peaks at an R^2 of 91,9%, indicating good model suitability to this regional dataset (Figure 4.8).

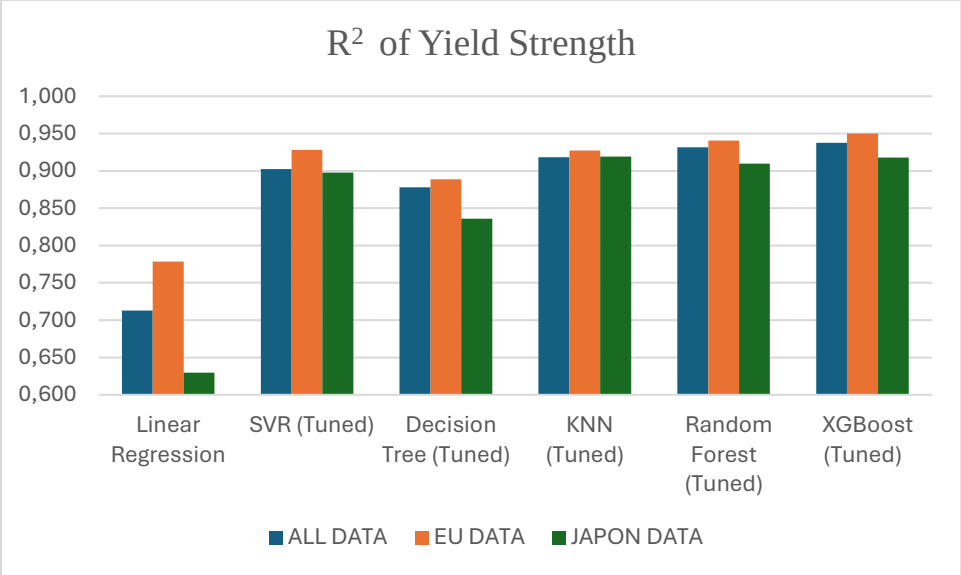

Figure 4.8 R^2 Performance of Machine Learning Models in Predicting Yield Strength Across Global, European, and Japanese Data Sets

Table 4.3 focuses on the impact of different feature selection methods on enhancing the R^2 values. For yield strength prediction using the complete dataset, the highest R^2 value of 94.1% was achieved when the 23rd feature was included through both Mutual Information and ANOVA F-Test feature selection methods (Figure 4.8, 4.10). When utilizing only the EU data, the LASSO method yielded the highest R^2 value at 95.4% (Figure 4.11). In the case of the Japanese data, the highest R^2 value of 94.5% was attained with the Sequential Feature Selection method when the 23rd feature was added. Figure 4.12 illustrates the results of applying Mutual Information Feature Selection to determine the effectiveness of 24 input features on Yield Strength predictions using the XGBoost

model. Features are added iteratively from the most to the least impactful based on their mutual information scores, showing how the R^2 value of the model's predictions evolves as each feature is included. The target line represents the benchmark R^2 value for optimal model performance.

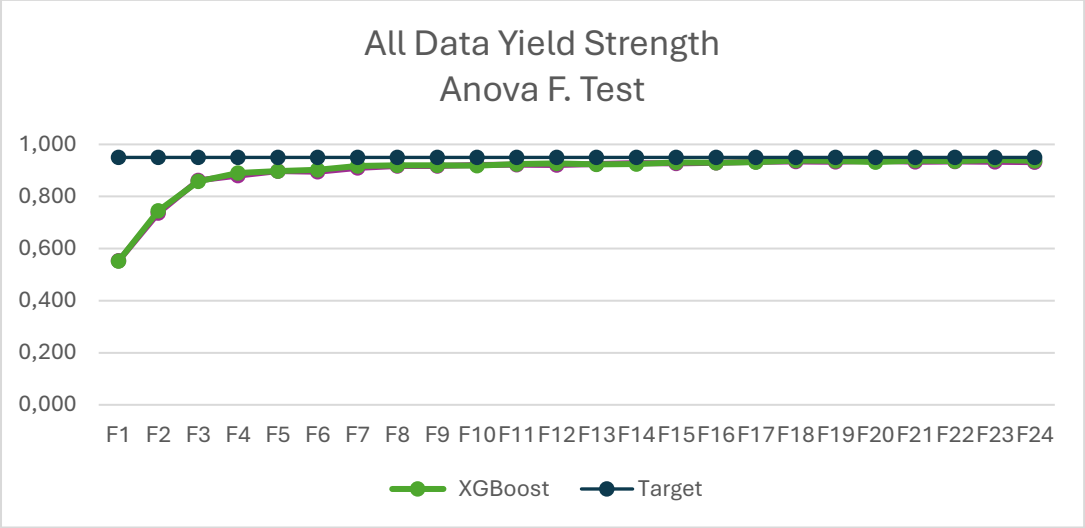

Figure 4.0.9 Yield Strength vs. Anova F-Test for All Data

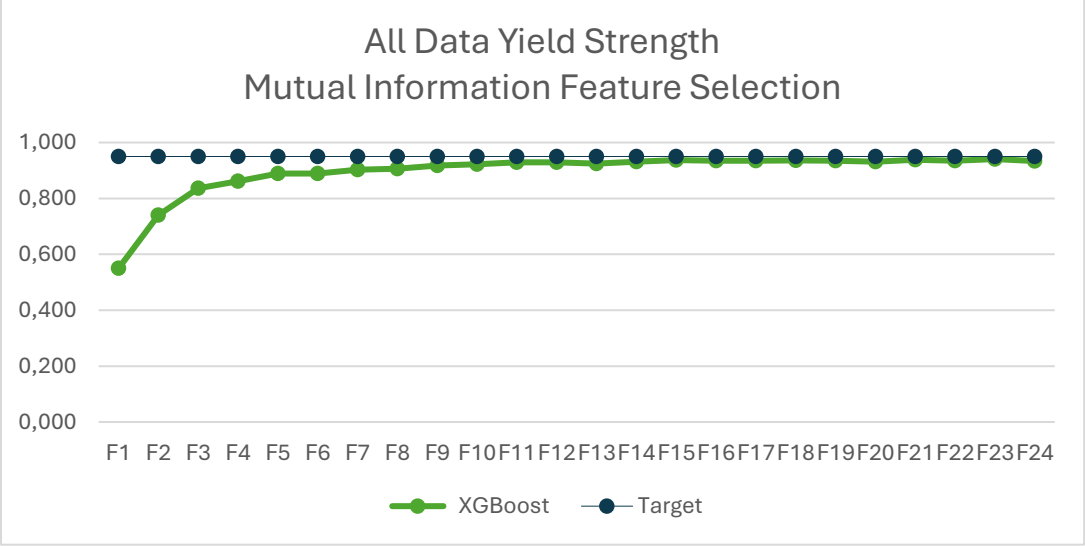

Figure 4.10 Yield Strength vs Mutual Information Feature Selection For All Data

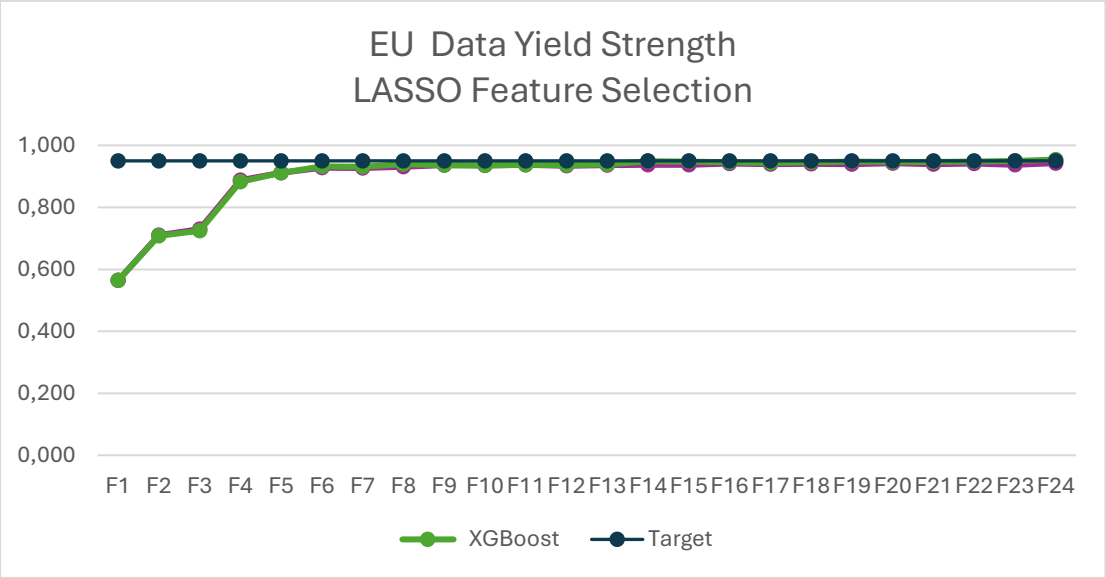

Figure 4.11 Yield Strength vs. Lasso for EU Data

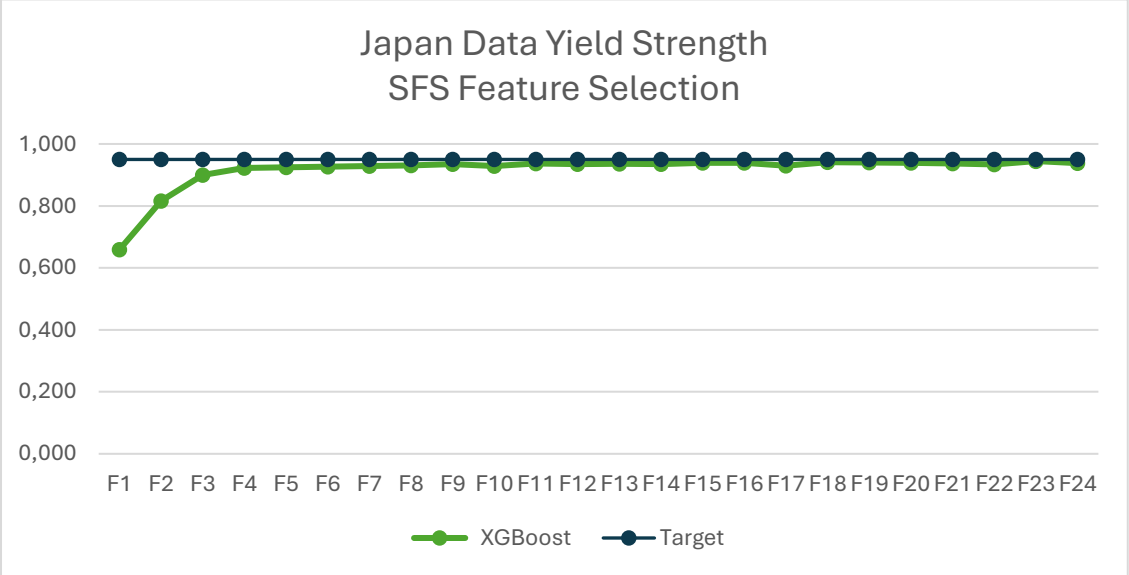

Figure 4.12 Yield Strength vs. SFS For Japan Data

For tensile strength, the models perform best when utilizing all available data, with the XGBoost model achieving the highest R^2 value of 90.8%. This suggests that a more comprehensive dataset provides a richer basis for the model to understand and predict material behaviors accurately. The EU data closely follows, with the KNN showing a slightly lower performance with an R^2 of 88.4%. On the other hand, the Japan data, while

slightly higher, still shows commendable results, especially with the KNN model, which peaks at an R^2 of 89.6%, indicating good model suitability to this regional dataset.

Considering the yield strength, this thesis, it is also exploring the efficacy of various feature selection methods applied to the XGBoost model across different regional datasets to predict steel properties. The key findings are listed below:

1. When employing the Mutual Information method for feature selection on European Union data, adding the 15th feature (Figure 4.13) iteratively into the XGBoost model yields an R^2 value of 95.3%. This indicates a high level of accuracy in predicting steel properties using this feature selection technique and dataset.

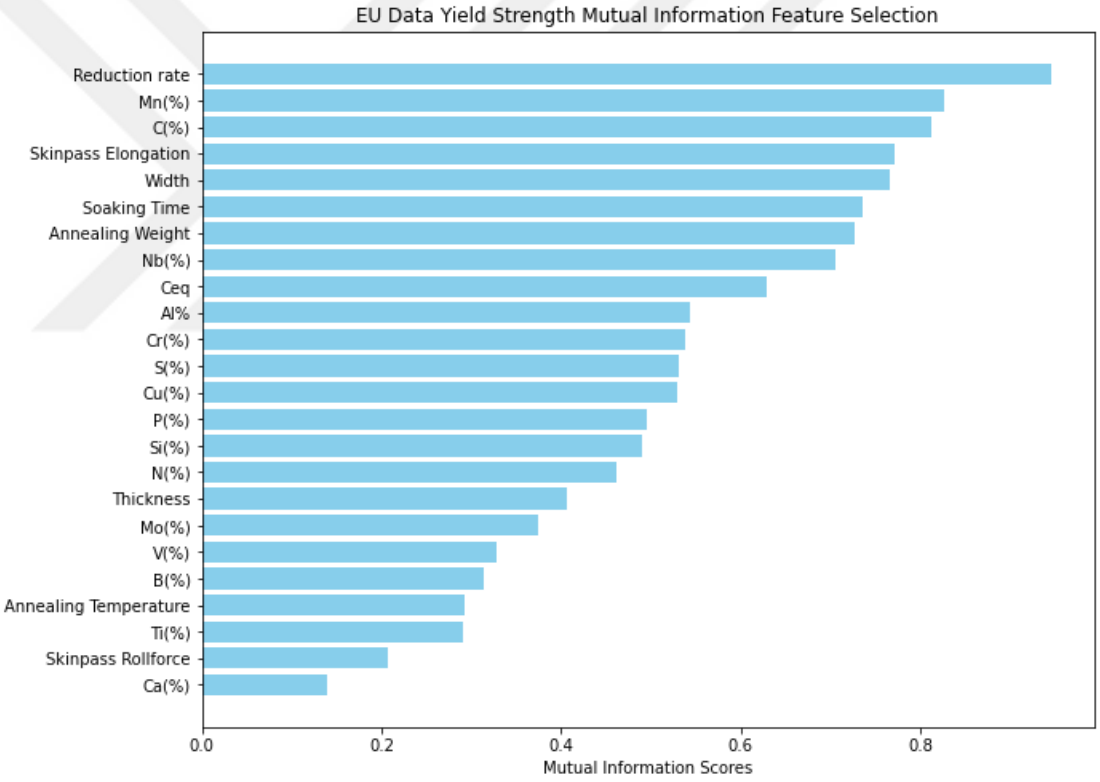

Figure 4.13 Mutual Feature Selection of Yield Strength of EU Data

2. Utilizing the Sequential Feature Selector across the entire dataset results in an R^2 value of 94% in the XGBoost model, iteratively including the 18th feature (Figure 4.14). This demonstrates that SFS is effective in optimizing the model's performance on a comprehensive data scope.

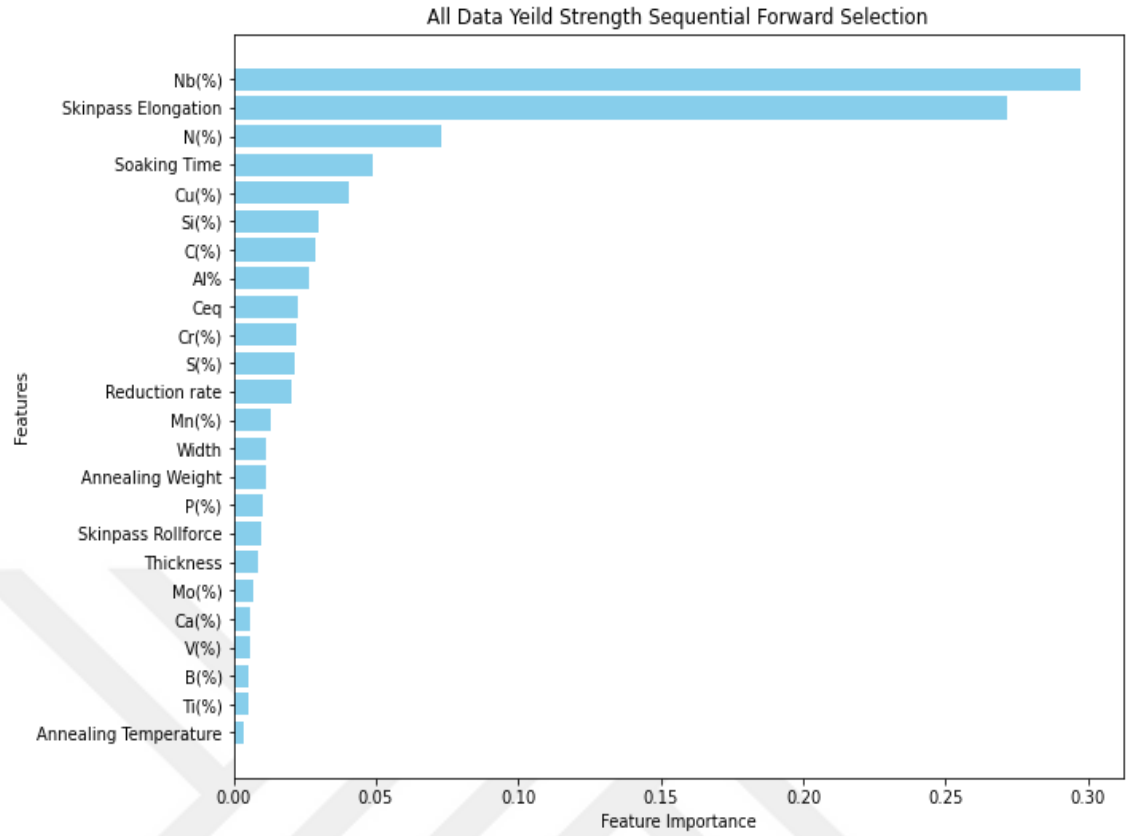

Figure 4.14 Sequential Forward Selection of Yield Strength of All Data

3. In the EU dataset, the application of the XGBoost-based feature selection process to iteratively include the 19th feature (Figure 4.15). as determined by its importance, yields an R^2 value of 95.3%. This highlights the robustness of the XGBoost algorithm's intrinsic feature ranking in enhancing model accuracy.

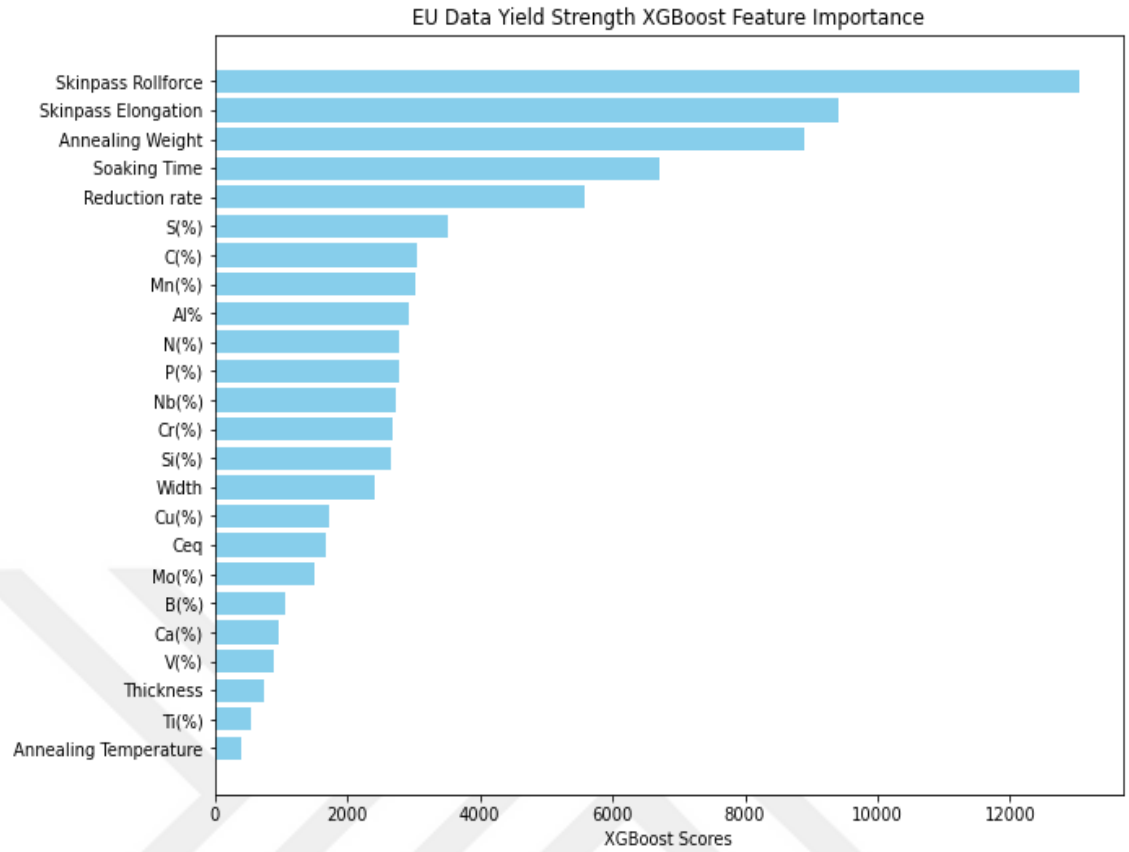

Figure 4.15 XGBoost Feature Importance of Yield Strength of EU Data

These findings substantiate the effectiveness of advanced feature selection methods in enhancing the predictive accuracy of machine learning models within the steel industry, particularly when regional variations in data are considered (Figure 4.16).

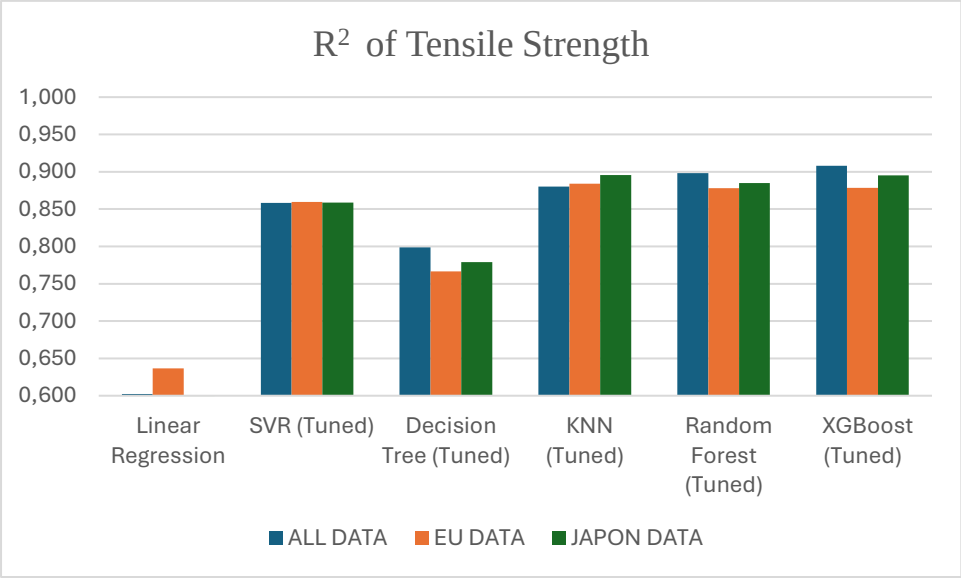

Figure 4.16 Performance Comparison of Machine Learning Models in Predicting Tensile Strength Across Global, European, and Japanese Data Sets

Considering the impact of different feature selection methods regarding tensile strength predictions, the highest R^2 value of 91.2% was obtained using the complete dataset with the inclusion of the 23rd feature through Mutual Information, ANOVA F-Test, and Recursive Feature Selection methods (Figure 4.17, Figure 4.18, Figure 4.19). For the EU dataset, the LASSO method provided the highest R^2 value of 88.6% with KNN model (Figure 4.20). The Japanese dataset showed an increase to a peak R^2 value of 91.6% when the 22nd feature was added using the SFS method (Figure 4.21).

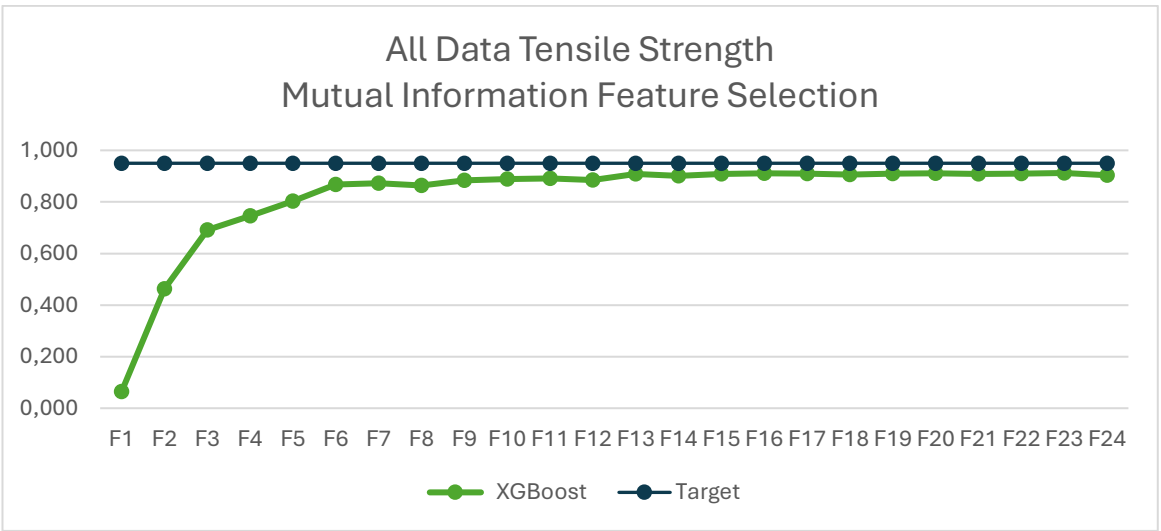

Figure 4.17 Tensile Strength vs. Mutual Information Feature Selection for All Data

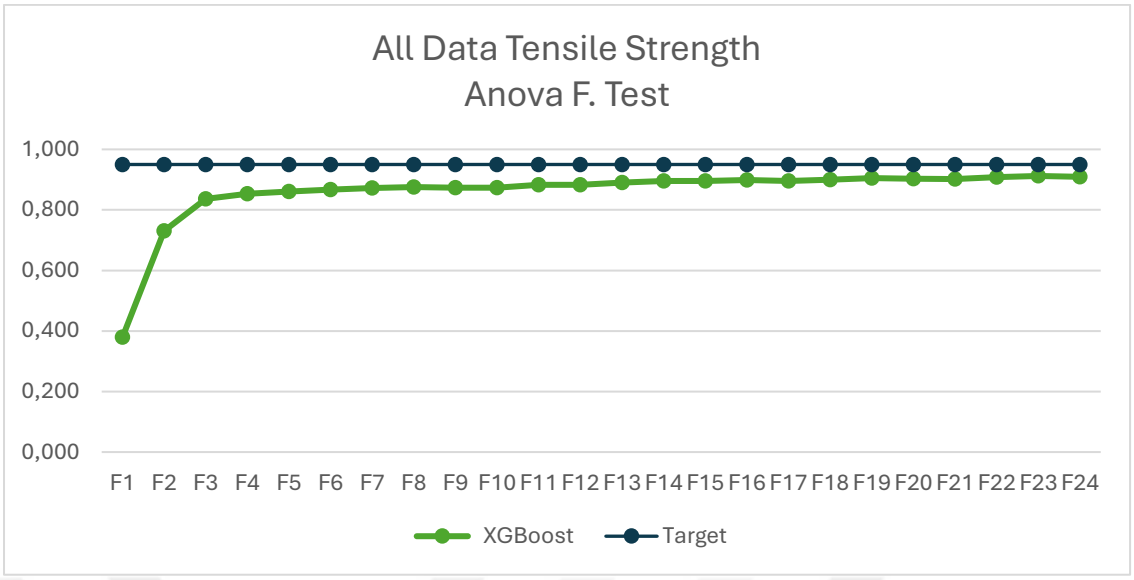

Figure 4.18 Tensile Strength vs Anova F-Test for All Data

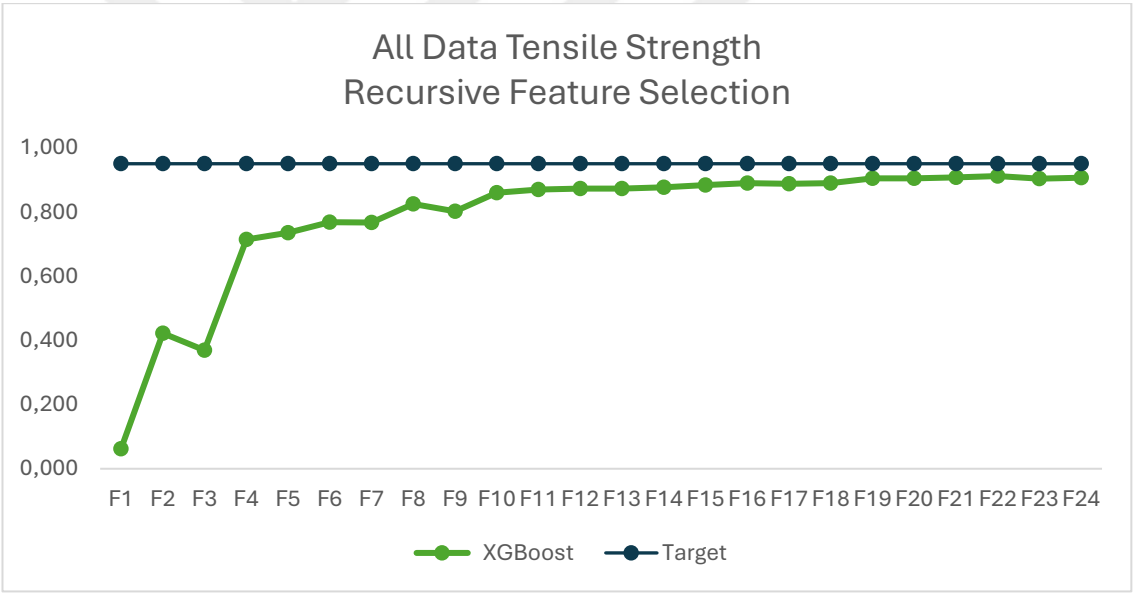

Figure 4.19 Tensile Strength vs. Recursive Feature Selection for All Data

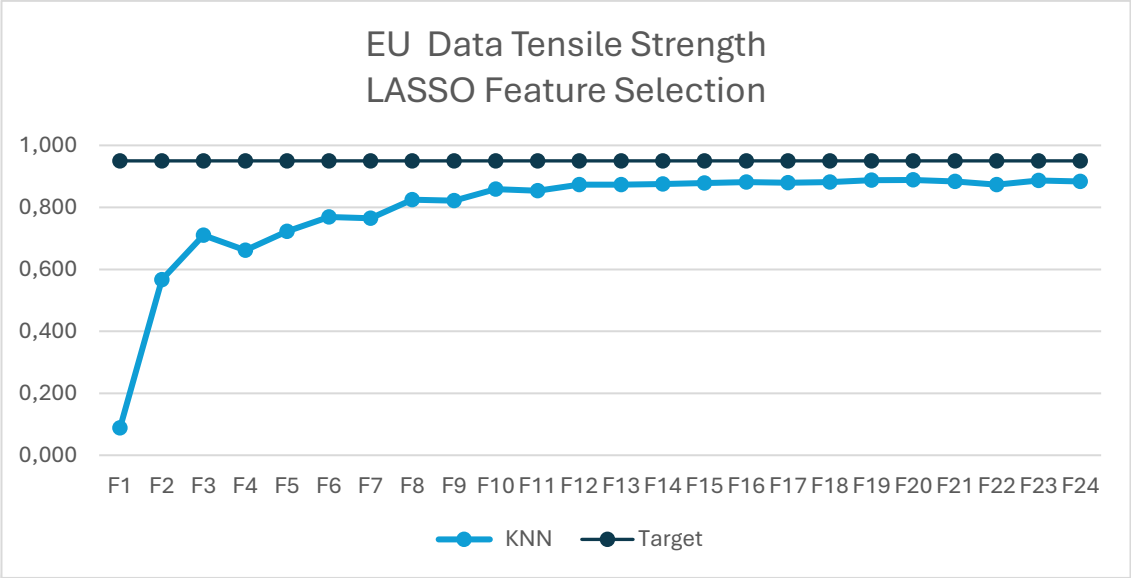

Figure 4.20 Tensile Strength vs Recursive Feature Selection for EU Data

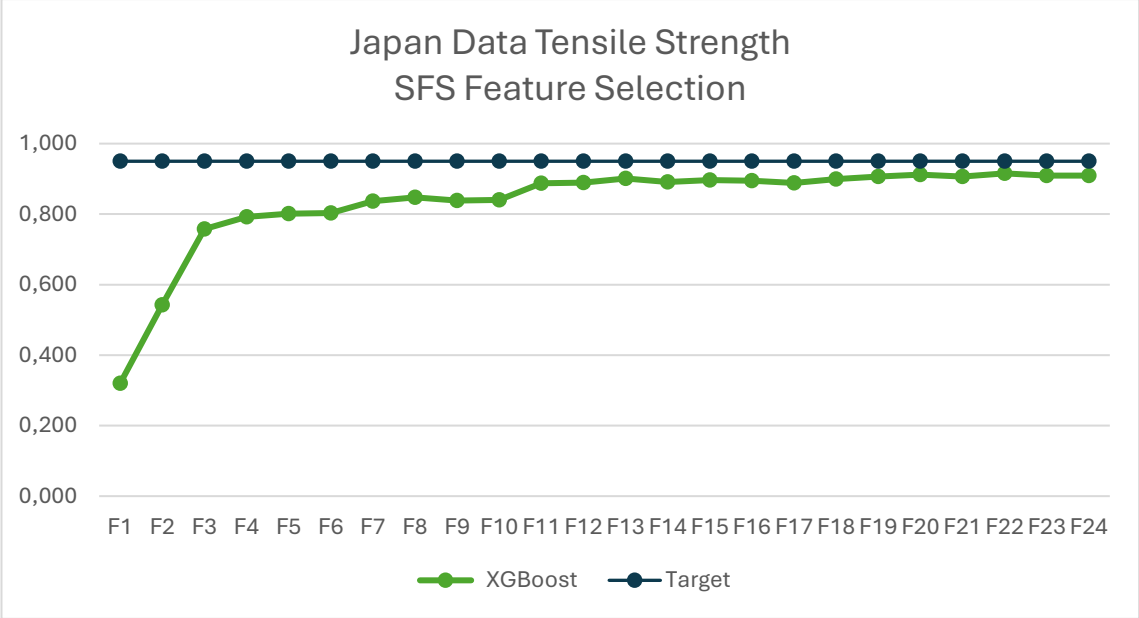

Figure 4.21 Tensile Strength vs SFS Selection for Japan Data

Moreover, these findings indicate that feature selection techniques significantly boost model performance, especially when customized to specific regional datasets, as demonstrated by the significant gains observed with the Japan data results.

Table 4.4 RMSE Performance of Machine Learning Models in Predicting Yield Strength Across Global, European, and Japanese Data Sets

	RMSE of Yield Strength			RMSE of Tensile Strength		
	All Data	EU Data	Japan Data	All Data	EU Data	Japan Data
Linear Regression	14,72	11,98	16,64	12,66	9,43	13
SVR	8,58	6,83	8,75	7,55	5,86	7,72
Decision Tree	9,95	8,07	11,07	8,97	7,9	9,65
KNN	7,86	6,87	7,78	6,95	5,33	6,64
Random Forest	7,32	6,08	8,22	6,46	5,61	6,97
XGBoost	6,93	5,58	7,83	6,09	5,49	6,65

Additionally, When the model-based RMSE values for yield and tensile strengths are analyzed, the following results are observed: for yield strength, a minimum RMSE of 6.933 is achieved with XGBoost when all available data are used, 5.588 with EU data, and 7.782 with Japan data using the KNN model (Table 4.4).

For tensile strength, a minimum RMSE of 69 is achieved with XGBoost when all data are used, 5.33 with EU data, and 6.64 with Japan data using the KNN model.

In the current study, R^2 values for yield and tensile strengths were found to range between 88.4 and 95 across various datasets. Root Mean Square Errors (RMSE) were also determined to be between 5 and 12 MPa for these datasets. According to the EN 10130 standard, a difference of 100 MPa is noted between the yield and tensile limits for the specified grades. Consequently, with R^2 values as indicators of model performance, an RMSE ranging from 5 to 12 MPa for the 5% deviation in data is considered to be at an acceptable level for real-world applications. This evaluation highlights the robustness of the models in predicting material properties within expected tolerances for industrial uses.

As demonstrated by the significant gains observed with the Japan data results by using SFS feature selection method.

In the context of yield strength, the strongest features identified were skin pass elongation (%) and annealing soaking time, from a metallurgical perspective, these parameters are critically influential in determining yield strength due to their direct impact on the microstructural properties of the metals (Figure 4.22).

The studies reveal that skin pass rolling, which involves a slight elongation of steel sheets, significantly impacts the yield strength of the steel. This process is employed to enhance surface quality and material properties. As skin pass elongation increases, there is an initial reduction in yield strength which reaches a minimum before increasing again with further elongation (Grassino et al. 2012). This behavior is attributed to the microstructural changes induced by the skin pass rolling. Specifically, the plastic strain introduced during the process leads to a more homogeneous deformation field across the grains exposed on the strip's surface.

Considering the correlation between annealing soaking time and yield strength in terms of a metallurgical point of view, as the soaking time increases during the annealing process at a temperature of 900 degrees Celsius, there is a notable decrease in yield strength, tensile strength and hardness (Raji and Oluwole 2012), which is particularly evident with a steep drop between 30 and 40 minutes of soaking time.

The reduction in yield strength with increased soaking time is attributed to changes in the steel's microstructure. During the annealing process, the metal undergoes recrystallization, where new grains form and grow, replacing the deformed structure created during cold drawing. This transformation leads to a reduction in dislocations, which are defects within the crystal structure that strengthen the metal by hindering the movement of atoms (Raji and Oluwole 2012). As these dislocations are reduced, the metal becomes less resistant to deformation, thus decreasing its yield strength. Additionally, the process relieves internal stresses built up during the initial cold drawing. This relaxation of stress contributes to the decrease in mechanical strength but improves

ductility, making the steel more malleable and better suited for further processing (Raji and Oluwole 2012). The literature highlights that optimal mechanical properties are achieved by balancing the effects of soaking time to achieve desired levels of yield strength and ductility tailored to specific industrial applications (Wichienrak and Puajindanetr 2019).

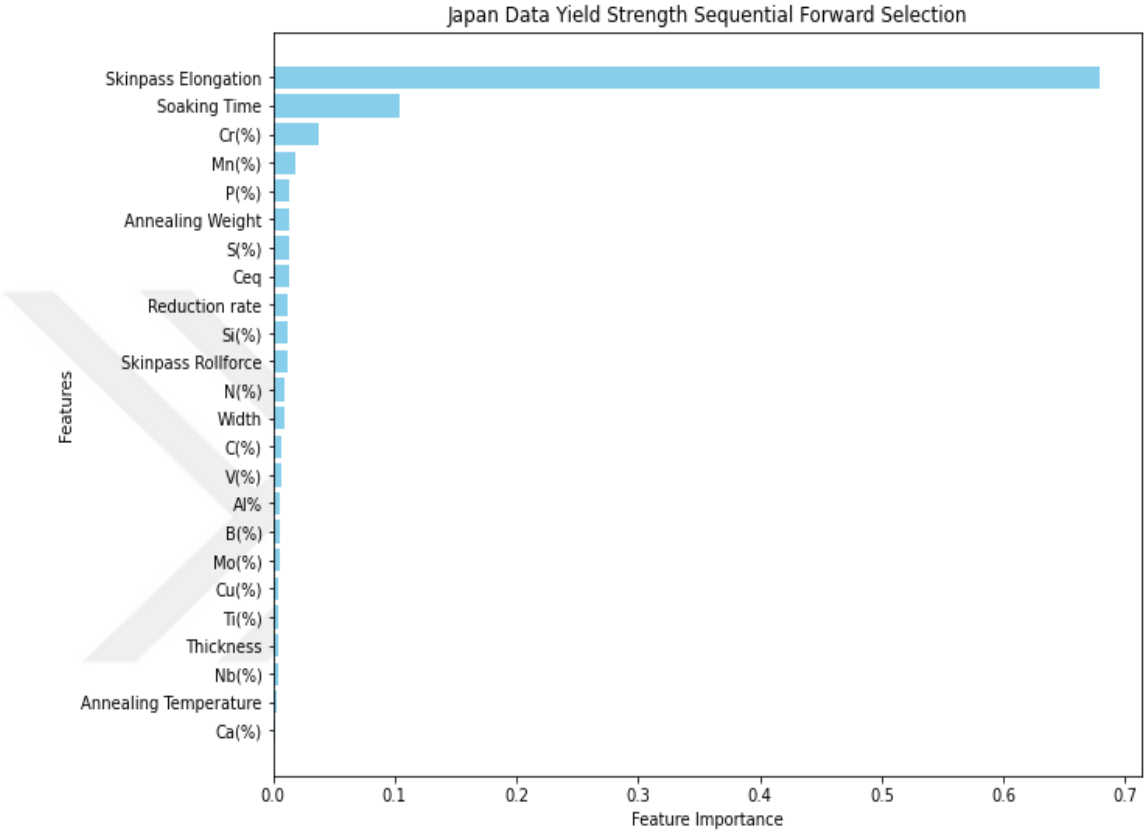

Figure 4.22 Feature Importance in Yield Strength Prediction for Japanese Data Using Sequential Forward Selection

For tensile strength, the most influential features were found to be the Niobium (Nb) content and Carbon Equivalent (Ceq) (Figure 4.23), measured as a percentage of the chemical composition. These elements play a pivotal role in enhancing the tensile strength of the material by modifying the grain structure and strengthening mechanisms within the metal matrix.

Niobium acts as a microalloying element that significantly enhances the tensile strength of the steel through grain refinement and precipitation strengthening mechanisms (Shanmugam et al. 2006).

The presence of niobium promotes the formation of fine-grained ferritic structures and complex carbides within the steel matrix. These microstructural changes are crucial because they increase the dislocation density within the material (Shanmugam et al. 2006). Higher dislocation densities impede the movement of dislocations, which is a primary mechanism of deformation in metals, thus enhancing the tensile strength of the steel. Moreover, niobium contributes to the formation of niobium carbides (NbC), which are effective at hindering grain growth during heat treatment processes such as annealing. By preventing grain growth, niobium ensures that the steel retains a fine-grained structure even after thermal cycles, which is essential for maintaining high strength and toughness (Shanmugam et al. 2006). The article points out that the precipitates observed are often of the MC type (metal carbides), where niobium carbides play a pivotal role in enhancing the mechanical properties by obstructing the dislocation motion more effectively (Shanmugam et al. 2006).

Regarding the correlation between carbon equivalent and tensile strength in terms of the metallurgical aspect, it is explained in the literature that increasing the carbon equivalent content in steels improves both the tensile strength and plasticity (C. Chen et al. 2022). The carbon plays a critical role by increasing the number of dislocations and enhancing the strain-hardening effect, which contributes to higher tensile strength.

These insights affirm the significant correlation between selected metallurgical features and the mechanical properties of the materials, illustrating the profound impact of precise chemical composition and processing conditions on the final product characteristics.

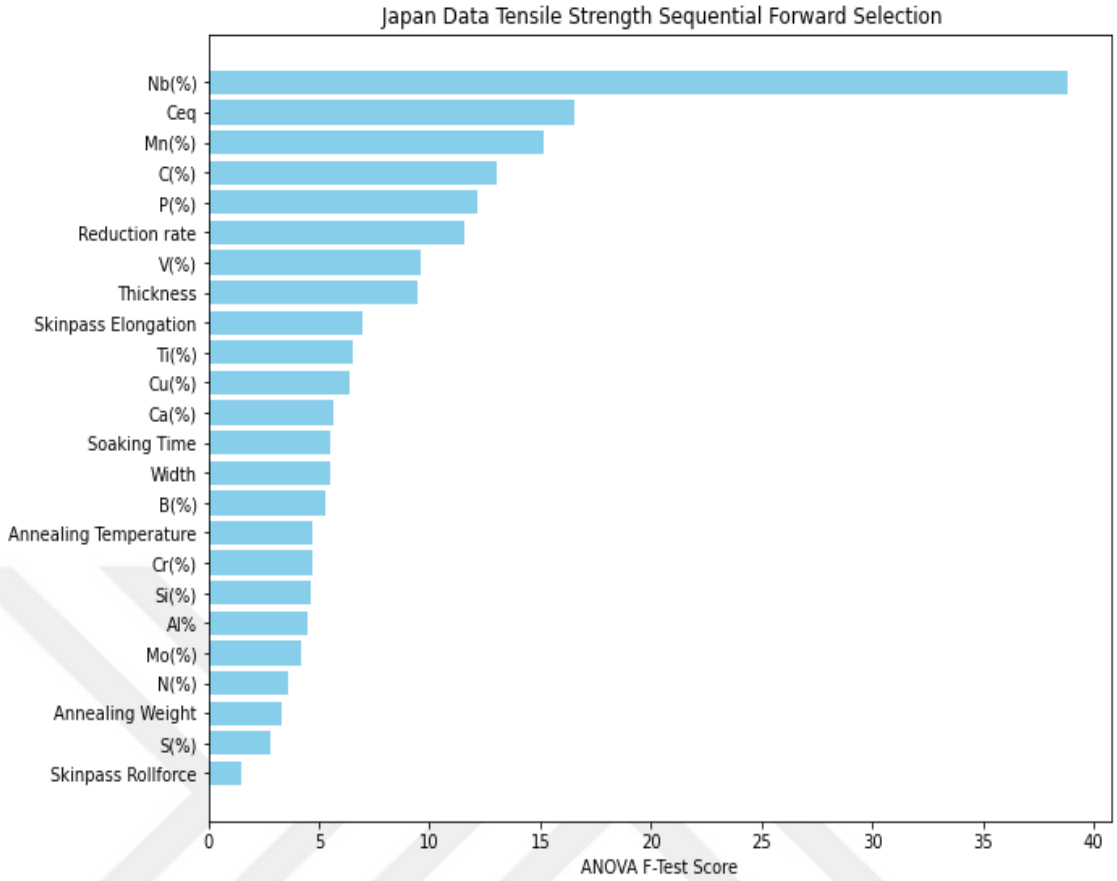

Figure 4.23 Analysis of Feature Importance for Tensile Strength Prediction in Japanese Data Using Sequential Forward Selection

5. CONCLUSION

This study involves using machine learning models Linear Regression Support Vector Regression, Decision Tree, K nearest neighbor, Random Forest, and XGBoost in conjunction with seven different feature selection methods (Recursive Feature Selection, Anova Feature Selection, Mutual Information Future Selection, XGBoost Feature Selection, Random Forest Feature selection) to predict mechanical properties of flat steel. The selected features were tested iteratively across models, and the outcomes of various feature selection methods were compared. The best performance results were then interpreted using domain knowledge.

These findings illustrate the effectiveness of employed feature selection methods in enhancing the predictive accuracy of mechanical property models across varied datasets. For the yield strength, using all data, the XGBoost model achieved an R^2 value of 93.8%. With the application of Mutual Information feature selection and the Anova F-Test, the R^2 value improved to 94.1% in the XGBoost model. For the EU data, the XGBoost model reached an R^2 value of 95%, which increased to 95.4% with the incorporation of LASSO feature selection. Utilizing Japan data, the initial R^2 with the XGBoost model was 91.9%, enhanced to 94.5% following SFS feature selection. Regarding the tensile strength results, when all data was employed, the XGBoost model yielded an R^2 of 90.8%, which was marginally improved to 91.2% through Mutual Information feature selection combined with the Anova F-Test. For EU data, the KNN model started with an R^2 of 88.4%, slightly rising to 88.6% after implementing Recursive feature selection. With Japan data, the KNN model initially had an R^2 of 89.6%, which substantially increased to 91.6% after the application of SFS feature selection and switching to the XGBoost model.

The findings of this thesis present strong evidence that the implementation of machine learning algorithms, with a particular emphasis on the XGBoost model, substantially improves the accuracy of predicting mechanical properties in cold-rolled flat steel coils. This advancement holds the potential to eliminate the necessity for physical test samples on the skin pass line, which is a significant step forward for the flat steel industry. By

utilizing these advanced predictive models, manufacturers can reduce scrap rates and lower production costs. The enhanced precision in predicting mechanical properties not only streamlines the production process but also contributes to greater efficiency and cost-effectiveness in the manufacturing of flat steel products. This study underscores the transformative impact of integrating machine learning techniques into industrial applications, highlighting the XGBoost model's capacity to drive significant operational improvements.

Several potential improvements and future directions have been identified for this study, which are crucial for advancing the field of predictive modeling in material science. Firstly, increasing the data size could significantly impact model performance. A larger dataset would provide more comprehensive information, potentially leading to more robust and accurate predictive models.

Secondly, feature extraction could also play a critical role in enhancing performance. Advanced feature extraction techniques could yield more relevant features, improving the models' predictive power and reliability. Future studies should prioritize exploring and refining these techniques.

Finally, the methodologies and findings of this study could be applied to different material qualities. Extending these techniques to various steel grades or other materials could validate the models' generalizability and effectiveness, highlighting their broader applicability.

Pursuing these future directions is essential for providing valuable insights and further advancing predictive modeling in material science.

In the final stage of processing cold-rolled flat steel products, specifically in the skin pass line, a sample for mechanical properties is collected to satisfy customer requirements. The development of a mechanical property prediction model offers two primary advantages: increased efficiency and cost savings. During the sampling phase, the final 8 meters of the coil are discarded, and a 500 mm wide sample is taken from the beginning

of the line's steady operation. This procedure typically consumes between 1 to 1.5 minutes per coil. A mechanical property test specimen is derived from this sample, followed by a destructive testing process. Implementing machine learning algorithms to predict mechanical properties can eliminate the need to remove eight meters of scrap from each coil. Furthermore, the time-consuming sampling process, which takes about 1 to 1.5 minutes per coil, would be unnecessary, thereby enhancing line productivity. The increased capacity from this improvement is estimated at roughly 0.5 tons per coil, marking a significant efficiency enhancement. Currently, the outputs of this study are utilized for the deployment of machine learning models within the manufacturing execution system, aiming to reduce the reliance on destructive testing. In the future, the findings of this thesis may be implemented across other production

BIBLIOGRAPHY

Ahmad, Ejaz, F Karim, K Saeed, Tanvir Manzoor, and G H Zahid. 2014. “Effect of Cold Rolling and Annealing on the Grain Refinement of Low Alloy Steel.” In *IOP Conference Series: Materials Science and Engineering*, 60:012029. IOP Publishing.

Akadi, Ali El, Abdeljalil El Ouardighi, and Driss Aboutajdine. 2008. “A Powerful Feature Selection Approach Based on Mutual Information.” *International Journal of Computer Science and Network Security* 8 (4): 116.

Amiri, Fatemeh, MohammadMahdi Rezaei Yousefi, Caro Lucas, Azadeh Shakery, and Nasser Yazdani. 2011. “Mutual Information-Based Feature Selection for Intrusion Detection Systems.” *Journal of Network and Computer Applications* 34 (4): 1184–99.

Battiti, Roberto. 1994. “Using Mutual Information for Selecting Features in Supervised Neural Net Learning.” *IEEE Transactions on Neural Networks* 5 (4): 537–50.

Beranger, Gerard, Guy Henry, and Germani Sanz. 1994. *The Book of Steel*. Paris cedex 08 France: Lavoisier Publishing.

Bhattacharyya, Tanmay, Shiv Brat Singh, Swati Sikdar Dey, Sandip Bhattacharyya, Wolfgang Bleck, and Debashish Bhattacharjee. 2013a. “Microstructural Prediction through Artificial Neural Network (ANN) for Development of Transformation Induced Plasticity (TRIP) Aided Steel.” *Materials Science and Engineering: A* 565:148–57.

———. 2013b. “Microstructural Prediction through Artificial Neural Network (ANN) for Development of Transformation Induced Plasticity (TRIP) Aided Steel.” *Materials Science and Engineering: A* 565:148–57.

Bradley, Paul S, and Olvi L Mangasarian. 1998. “Feature Selection via Concave Minimization and Support Vector Machines.” In *ICML*, 98:82–90.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45:5–32.

Cawley, Gavin, Nicola Talbot, and Mark Girolami. 2006. “Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation.” *Advances in Neural Information Processing Systems* 19.

Chandrashekar, Girish, and Ferat Sahin. 2014. “A Survey on Feature Selection Methods.” *Computers & Electrical Engineering* 40 (1): 16–28.

Chemmakha, Mohammed, Omar Habibi, and Mohamed Lazaar. 2022. “Improving Machine Learning Models for Malware Detection Using Embedded Feature Selection Method.” *IFAC-PapersOnLine* 55 (12): 771–76.

Chen, Chen, Hua Ma, Fei Wang, Zhinan Yang, Fucheng Zhang, and Zehui Yan. 2022. “Influence of Carbon Content on Tensile Properties of Pure High Manganese Austenitic Steel.” *Coatings* 12 (11): 1622.

Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Chen, Xue-wen, and Jong Cheol Jeong. 2007. “Enhanced Recursive Feature Elimination.” In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 429–35. IEEE.

Chou, JS, NT Ngo, WK Chong - Engineering Applications of Artificial, and undefined 2017. n.d. “The Use of Artificial Intelligence Combiners for Modeling Steel Pitting Risk and Corrosion Rate.” *ElsevierJS Chou, NT Ngo, WK ChongEngineering*

Applications of Artificial Intelligence, 2017•Elsevier. Accessed April 18, 2024. https://www.sciencedirect.com/science/article/pii/S0952197616301737.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97. https://doi.org/10.1007/BF00994018.

Costa, Ana P O, Mariana R R Seabra, José M A César de Sá, and Abel D Santos. 2024. "Manufacturing Process Encoding through Natural Language Processing for Prediction of Material Properties." *Computational Materials Science* 237:112896.

Costa, APO, MRR Seabra, … JMAC de Sá - Computational Materials, and undefined 2024. n.d. "Manufacturing Process Encoding through Natural Language Processing for Prediction of Material Properties." *ElsevierAPO Costa, MRR Seabra, JMAC de Sá, AD SantosComputational Materials Science, 2024•Elsevier*. Accessed March 30, 2024. https://www.sciencedirect.com/science/article/pii/S0927025624001174.

E. Dowling Norman, Frank Maher, Siva Prasad Katakam, and R. Narayansamy. 2013. *Mechanical Behavior of Material*. Edited by Holly: Executive Editor Stark, Disanno Scott : Senior Managing Editor, and Sandin Daniel: Media Editor. Fourth Edition.

Elssied, Nadir Omer Fadl, Othman Ibrahim, and Ahmed Hamza Osman. 2014. "A Novel Feature Selection Based on One-Way Anova f-Test for e-Mail Spam Classification." *Research Journal of Applied Sciences, Engineering and Technology* 7 (3): 625–38.

Fonti, Valeria, and Eduard Belitser. 2017. "Feature Selection Using Lasso." *VU Amsterdam Research Paper in Business Analytics* 30:1–25.

Géron, Aurélien. 2022. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Grassino, Jacopo, Maurizio Vedani, Gianmarco Vimercati, and Guido Zanella. 2012. "Effects of Skin Pass Rolling Parameters on Mechanical Properties of Steels." *International Journal of Precision Engineering and Manufacturing* 13:2017–26.

Hackeling, G. 2017. "Mastering Machine Learning with Scikit-Learn." https://books.google.com/books?hl=tr&lr=&id=9-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=Scikit-learn+documentation+on+StandardScaler.&ots=FOaAPx9XOi&sig=C_PVLCdNNuRBWHZmlli-TPPkk-U.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.

Huljanah, Mia, Zuherman Rustam, Suarsih Utama, and Titin Siswantining. 2019. "Feature Selection Using Random Forest Classifier for Predicting Prostate Cancer." In *IOP Conference Series: Materials Science and Engineering*, 546:052031. IOP Publishing.

Jiang, Zheyong, Jinxing Che, Mingjun He, and Fang Yuan. 2023. "A CGRU Multi-Step Wind Speed Forecasting Model Based on Multi-Label Specific XGBoost Feature Selection and Secondary Decomposition." *Renewable Energy* 203:802–27.

Jović, Alan, Karla Brkić, and Nikola Bogunović. 2015. "A Review of Feature Selection Methods with Applications." In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. Ieee.

König, Gunnar, Christoph Molnar, Bernd Bischl, and Moritz Grosse-Wentrup. 2021. "Relative Feature Importance." In *2020 25th International Conference on Pattern Recognition (ICPR)*, 9318–25. IEEE.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “Imagenet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 25.

Lee, Jinlee, Dooho Park, and Changhoon Lee. 2017. “Feature Selection Algorithm for Intrusions Detection System Using Sequential Forward Search and Random Forest Classifier.” *KSII Transactions on Internet and Information Systems (TIIS)* 11 (10): 5132–48.

Li, Jie, Lanjia Pan, Manu Suvarna, Yen Wah Tong, and Xiaonan Wang. 2020. “Fuel Properties of Hydrochar and Pyrochar: Prediction and Exploration with Machine Learning.” *Applied Energy* 269:115166.

Lugan, A, PA Hilton, … DW Taylor - Congress on Applications of Lasers &, and undefined 2002. n.d. “The Effects of Steel Composition on the Laser Cut Edge Quality of Carbon and C-Mn Steels.” *Pubs.Aip.Org*. Accessed March 30, 2024. https://pubs.aip.org/liacp/proceedings-abstract/ICALEO/2002/893119.

Ma, Shuangge, and Jian Huang. 2008. “Penalized Feature Selection and Classification in Bioinformatics.” *Briefings in Bioinformatics* 9 (5): 392–403.

Maldonado, Sebastián, Richard Weber, and Fazel Famili. 2014. “Feature Selection for High-Dimensional Class-Imbalanced Data Sets Using Support Vector Machines.” *Information Sciences* 286:228–46.

Morales-España, G, Juan Mora-Flórez, and Gilberto Carrillo-Caicedo. 2010. “A Complete Fault Location Formulation for Distribution Systems Using the K-Nearest Neighbors for Regression and Classification.” In *2010 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America (T&D-LA)*, 810–15. IEEE.

Nguyen, Bac, Carlos Morell, and Bernard De Baets. 2016. “Large-Scale Distance Metric Learning for k-Nearest Neighbors Regression.” *Neurocomputing* 214:805–14.

Pal, Mahesh, N K Singh, and N K Tiwari. 2011. “Support Vector Regression Based Modeling of Pier Scour Using Field Data.” *Engineering Applications of Artificial Intelligence* 24 (5): 911–16.

Pekel, Engin. 2020. “Estimation of Soil Moisture Using Decision Tree Regression.” *Theoretical and Applied Climatology* 139 (3): 1111–19.

Raji, Nurudeen Adekunle, and Oluleke Olugbemiga Oluwole. 2012. “Effect of Soaking Time on the Machinical Properties of Anealed Cold-Drawn Low Carbon Steel.”

Reddy, ACS, S Rajesham, … PR Reddy - AIP Conference, and undefined 2020. 2020. “Formability: A Review on Different Sheet Metal Tests for Formability.” *Pubs.Aip.Org*. https://doi.org/10.1063/5.0019536.

Ridzuan, Fakhitah, and Wan Mohd Nazmee Wan Zainon. 2019. “A Review on Data Cleansing Methods for Big Data.” *Procedia Computer Science* 161:731–38.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1986. “Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, Ed. DE Rumelhart and J. McClelland. Vol. 1. 1986.” *Biometrika* 71:599–607.

Sánchez-Maroño, Noelia, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. 2007. “Filter Methods for Feature Selection–a Comparative Study.” In *International Conference on Intelligent Data Engineering and Automated Learning*, 178–87. Springer.

Sandri, Marco, and Paola Zuccolotto. 2006. “Variable Selection Using Random Forests.” In *Data Analysis, Classification and the Forward Search: Proceedings of the*

Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005, 263–70. Springer.

Sethi, Ishwar K. 1997. "Structure-Driven Induction of Decision Tree Classifiers through Neural Learning." *Pattern Recognition* 30 (11): 1893–1904.

Shanmugam, S, R D K Misra, J Hartmann, and S G Jansto. 2006. "Microstructure of High Strength Niobium-Containing Pipeline Steel." *Materials Science and Engineering: A* 441 (1–2): 215–29.

Su, Xiaogang, Xin Yan, and Chih Ling Tsai. 2012. "Linear Regression." *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (3): 275–94. https://doi.org/10.1002/WICS.1198.

Tang, Bo, and Haibo He. 2015. "ENN: Extended Nearest Neighbor Method for Pattern Recognition [Research Frontier]." *IEEE Computational Intelligence Magazine* 10 (3): 52–60.

Technology, H Kijima - Journal of Materials Processing, and undefined 2015. n.d. "An Experimental Investigation on the Influence of Lubrication on Roughness Transfer in Skin-Pass Rolling of Steel Strip." *ElsevierH KijimaJournal of Materials Processing Technology, 2015•Elsevier*. Accessed March 30, 2024. https://www.sciencedirect.com/science/article/pii/S0924013615002277.

Wichienrak, Ruangyot, and Somchai Puajindanetr. 2019. "Factors Affecting the Mechanical Properties Variation after Annealing of Cold Rolled Steel Sheet." In *E3S Web of Conferences*, 95:04003. EDP Sciences.

Xie, Q, M Suvarna, J Li, X Zhu, J Cai, X Wang - Materials & Design, and undefined 2021. n.d. "Online Prediction of Mechanical Properties of Hot Rolled Steel Plate Using Machine Learning." *ElsevierQ Xie, M Suvarna, J Li, X Zhu, J Cai, X WangMaterials & Design, 2021•Elsevier*. Accessed March 24, 2024. https://www.sciencedirect.com/science/article/pii/S026412752030736X.

Xie, Qian, Manu Suvarna, Jiali Li, Xinzhe Zhu, Jiajia Cai, and Xiaonan Wang. 2021. "Online Prediction of Mechanical Properties of Hot Rolled Steel Plate Using Machine Learning." *Materials & Design* 197:109201.

Xu, Min, Pakorn Watanachaturaporn, Pramod K Varshney, and Manoj K Arora. 2005. "Decision Tree Regression for Soft Classification of Remote Sensing Data." *Remote Sensing of Environment* 97 (3): 322–36.

Xu, Zhi-Wei, Xiao-Ming Liu, and Kai Zhang. 2019. "Mechanical Properties Prediction for Hot Rolled Alloy Steel Using Convolutional Neural Network." *Ieee Access* 7:47068–78.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–20.

CURRICULUM VITAE

Personal Information
Name and surname: Didem Bakiler İlme

Academic Background
Middle East Technical University
Metallurgical and Material Engineering
Foreign Languages: English

Publications and Presentations Derived from the Thesis
IEEE-UBMK-2024, 9th International Conference on Computer Science and Engineering, Comparison of Feature Selection Methods for Mechanical Properties of Cold Rolled Products in Flat Steel Manufacturing
(Authors: Didem Bakiler İlme, Merve Öper, E.Fatih Yetkin)

Work Experience
Metallurgy Quality Engineer at Borcelik between 2004 - 2011
Metallurgy Quality Division Manager at Borcelik between 2011-2017
Metallurgy and Quality Manager at Borcelik since 2017