Sigma J Eng Nat Sci, Vol. 43, No. 4, pp. 1088–1099, August, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00107

Research Article

# Application of data analytics and machine learning: A steel manufacturing facilities

**Ezgi TOSUN[1,*], Ali Fuat GUNERI[1]**

[1]*Department of Industrial Engineering, Yıldız Technical University, Istanbul, 34220, Türkiye*

## ARTICLE INFO

*Article history*
Received: 01 April 2024
Revised: 01 July 2024
Accepted: 03 August 2024

**Keywords:**
High-Quality Steel, Rolling Mill, Scrap Quantity, Data Set, Data Analysis, Machine Learning

## ABSTRACT

Statistically, global steel production data, especially regarding specialty steel, is crucial in high-quality steel manufacturing. This study focused on estimating the scrap generated during the production of round steel material in the rolling mill of a high-quality steel manufacturing facility. The impact of element ratios in the steel on scrap quantity was analyzed. A dataset was created using Oracle PL/SQL and Python, which was then cleansed of outliers. The analysis identified specific quality and dimensions linked to the highest scrap quantities, as well as the quality most responsible for scrap based on production input. The dimensions and quality associated with the highest production volumes were also determined. The element ratios within the material were examined to ascertain which element significantly influenced the scrap quantity. Additionally, it was analyzed which quality and size consumed more energy, impacting material pricing. Seven machine learning algorithms were developed, including four regression and three classification algorithms. These algorithms were evaluated using performance metrics. Among the regression algorithms, the Random Forest algorithm showed the best overall performance. For the classification algorithms, the K-Nearest Neighbors algorithm exhibited the best overall performance. In addition, an application was developed to display the results of model performance metrics based on the input parametric values.

**Cite this article as:** Tosun E, Guneri AF. Application of data analytics and machine learning: A steel manufacturing facilities. Sigma J Eng Nat Sci 2025;43(4):1088–1099.

## INTRODUCTION

In statistical terms, when looking at the 2023 data for global steel production, a facilities country is among the top 10 producers of high-quality steel [1]. Steel in all its forms is ubiquitous in our lives. High-quality steels, distinct from other types of steel, offer durability, efficiency, and safety in the specific areas where they are utilized [2]. The product group represented by high-quality steel finds applications in numerous fields, primarily emerging as a critical material in the defense industry or automotive sector. Due to its lightweight nature, durability, ease of inspection, interchangeability, and recyclability, specialty steel is utilized in many other sectors as well [3]. Within the scope of this study, data analytics and machine learning algorithms will

***Corresponding author.**
*E-mail address: ezgi.tosun@std.yildiz.edu.tr

*This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic*

Published by Yıldız Technical University Press, İstanbul, Turkey

be employed to estimate the amount of scrap generated in the of round steel material in the rolling mill production area of a company specializing in producing high-quality steel as raw material. Additionally, the analysis will examine the impact of elements contained within the high-quality steel material on the scrap quantity. The element whose impact on scrap quantity is greater will be identified. Additionally, different results of data analytics will also be included in the study an designed a basic application.

**Literatur Review**

Ziqiu K. and their friends conducted a literature review of applications carried out in the manufacturing facility using machine learning algorithms. In their study, they indicated that the application of machine learning algorithms for analyzing production lines aimed to improve quality control, assess risks, and additionally, reduce costs. In their study, they examined 271 articles, and conducted a detailed analysis on 39 foundational studies. During the literature review, they identified production line issues, targeted sectors, and analyzed which machine learning algorithms were utilized in the studies. They conducted a literature review focusing on Overall Equipment Effectiveness (OEE) estimation as their main topic. In their study, they concluded that chemical production and the metal industry were the most prominent sectors examined. They found that supervised machine learning algorithms were more commonly used as algorithms in their study. They found that approximately 60% of the selected articles for analysis were focused on regression models. In their study, they determined that artificial neural networks algorithm was the most commonly used algorithm. After artificial neural networks, they mentioned that support vector machines, k-nearest neighbours, gradient boosted decision trees, convolutional neural networks, and long short-term memory were among the most commonly used algorithms. They stated that the abundance of data in production lines encourages the use of machine learning algorithms [4].

Koruyan conducted a study on the classification of customer complaints using machine learning algorithms. In this study, they defined four different categories and created models using logistic regression, support vector machines, and probabilistic gradient descent algorithms. Following the establishment of the models, they conducted model performance evaluations and determined that logistic regression exhibited the best performance [5].

Mirasçı S. and their friend conducted a study utilizing both data mining and machine learning algorithms in the steel material procurement process. In this study, which began with the aim of making strategic procurement decisions for steel materials and increasing company profitability, they prepared the dataset for models using data mining. Subsequently, they created algorithms using hierarchical clustering and the K-means method. In this analysis, the study was conducted by determining the ideal number of clusters. Following this study, the suggested analyses helped reduce the negative impact of procurement department personnel on the procurement process strategy in the steel material procurement process. The study indicated that the analysis, which would otherwise take a long time, was achieved more quickly with machine learning algorithms [6].

Qian X. and their friends conducted a study on predicting the mechanical properties of steel plates using deep neural network algorithms. In their study, they mentioned the difficulty in determining the relationship between process parameters and mechanical properties. They mentioned that in their study, the deep neural networks algorithm achieved an R2 value of 90%. They also noted that the model performance was better compared to other classical machine learning algorithms [7].

Sandhya N. and their friends conducted a study using data science techniques to predict the mechanical properties of stainless steel. In their study, they aimed to use different algorithms to predict the mechanical properties of stainless steel. They utilized random forest algorithm, artificial neural networks, K-nearest neighbour, decision tree, support vector algorithm, linear regression, and batch method algorithm. In their study on stainless steel, they stated that the random forest algorithm showed the best performance in predicting the tensile strength of steel based on parameters such as carbon content and cross-section dimensions, while the K-nearest neighbour algorithm exhibited the worst performance as their conclusion [8].

S. M. Taslim Uddin Raju and their friends, their research introduces a robust framework tailored for forecasting demand within the steel industry. It encompasses a series of steps including data preprocessing, transformation, and feature selection, followed by the application of ensemble models such as bagging, boosting, and stacking. Alongside ensemble methods, the study also incorporates various machine learning approaches including Support Vector Regression, extreme learning machine and Multilayer Perceptron Neural Network as benchmarks. Through meticulous hyperparameter tuning via grid search, the framework aims to optimize model performance metrics such as the determination coefficient and root mean square error. Utilizing a consistent experimental setup with a steel industry dataset, the study evaluates the efficacy of different models Findings highlight STACK1 (ELM + GBR + XGBR-SVR) and STACK2 (ELM + GBR + XGBR-LASSO) models as superior performers, demonstrating the highest accuracies with $R^2$ values reaching 0.97 [9].

Chahbi I. and their friends, their article presents a new machine learning approach for predicting energy consumption in industrial settings. The proposed approach was applied using the Random Forest algorithm to predict energy consumption in the steel industry. The reason for selecting the Random Forest algorithm over other algorithms is noted to be its ability to provide the most effective prediction results and the importance of permutation feature, which helps steel industry experts better understand

the model's judgments. Consequently, they stated that it is an algorithm that will assist in optimizing energy consumption in industrial settings and making the most accurate decisions [10].

Choi S. and their colleagues conducted a study in response to the need for the steel industry to transition from traditional blast furnaces to electric arc furnaces to reduce carbon emissions, while lacking confidence in operator proficiency to determine the electrical power input for the electric arc furnace. Employing a data-driven approach, they developed a model using the support vector algorithm to enable real-time prediction of tap temperature, along with estimation of the input power quantity in the production area to enhance efficiency. At the conclusion of this study, they found that compared to operation based on operator discretion, tap temperature deviation decreased by 17%, and average power consumption decreased by up to 282 kWh/ton of heat, thus demonstrating significant improvements in efficiency [11].

Shiraiwa T. and their colleagues developed a machine learning model to predict fatigue strength with high accuracy. In their study, they utilized hierarchical clustering, linear regression, and artificial neural networks algorithms. They first selected a group of carbon steels using hierarchical clustering, then performed predictions using linear regression and artificial neural networks, and found that the prediction results were consistent with real data. [12].

Patel S.V. and their colleagues applied machine learning algorithms to diagnose temperature variation defects in the production of soft steel rolls. In soft steel production, they accessed data using sensors placed in the production area and after determining the relevant parameters, they constructed their models. They built their models using decision trees, artificial neural networks, support vector machines, random forests, and ensemble methods algorithms, and noted that the random forest algorithm performed the best among them [13].

Ecemis O. and their colleagues applied machine learning algorithms to predict sales based on the sectors served by a company in the stainless steel industry. After preprocessing the data, they created models using support vector machines and artificial neural networks. At the conclusion of their study, they noted that support vector machines exhibited better model performance [14].

**Data Set**

The dataset has been obtained from the company's system. The data pertains to the production area of the rolling mill where the application will be implemented. The data set consists of 29 columns and 2560 rows. As seen in Table 1, the order no is special number each order. The quality name is the quality information associated with the order and the round material. The diameter is diameter of the round material, e.g. A diameter. The billet size is billet size of the round material, e.g. AxA, BxB, CxC. The casting no corresponds to any of the castings associated with each order and special number each casting. The material ınput amount and material output amount is the amount of scrap belong production area, e.g. 1000 kilos. The ratio of elements are percentages of elements present within the round materail, e.g. element A in X ratio. In the study, it was assumed that the rolling speed and temperature in the production area were constant. Additionally, some variable names have been obscured for data privacy. In table 1 below, the columns of the data set are shown as an example.

## MATERIALS AND METHODS

The process flow diagram of the overall study is as shown in Figure 1.

**Table 1.** Study data set introduction

| | | | | | | | | Element Ratio | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Order No** | **Quality Name** | **Diameter** | **Billet Size** | **Casting No** | **Material Input Amount** | **Material Output Amount** | **Scrap Amount** | **A** | **B** | **C** | **D** | **C** | **E** | **F** | **G** | **H** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **R** | **S** |
| TTT | YYY | A | A | ZZZ | S | U | X | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

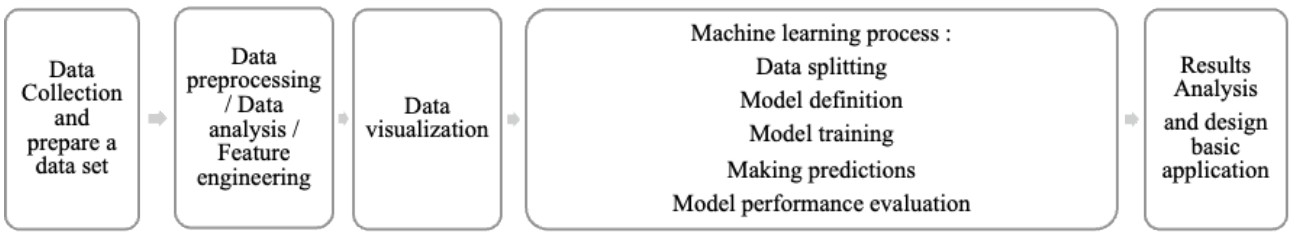


**Figure 1.** The study flow chart.

The study began with the creation of the dataset, followed by data analysis, and then predictions were made using machine learning algorithms, with the evaluation of model performance. A dataset is a collection of data. A dataset can consist of one or more tables, where each column and each row represents a data record. [15]. Data sets can be created from a database, as in this study. Data analysis is the process of examining, cleaning, transforming, and modeling data with the aim of exploring information, interpreting results, and supporting decision-making [16]. Machine Learning is a subfield of artificial intelligence that enables a system to learn from existing data without explicit programming, thereby enhancing its adaptation to new data and facilitating learning [17]. Machine learning algorithms are used for prediction or classification tasks. Given labeled or unlabeled input data, they will generate predictions about a model [18].

In this study, Oracle PL/SQL and Python programming languages will be used. The dataset related to the production area of the company's rolling mill has been created from the database using Oracle PL/SQL programming language for the application of machine learning algorithms. Some data analyses were also conducted additionally in Oracle PL/SQL. To prevent potential underfitting or overfitting situations in the dataset transferred to the Python programming language, outlier analysis was conducted, and outlier data were removed from the dataset. As a result of this process, machine learning algorithms exhibited better performance.

In Python programming language, visualization and machine learning algorithms were created using the pandas, numpy, matplotlib, seaborn, and scikit-learn libraries [19]. The data visualization screenshot is in below, figure 2, figure 3, figure 4, figure 5 and figure 6.

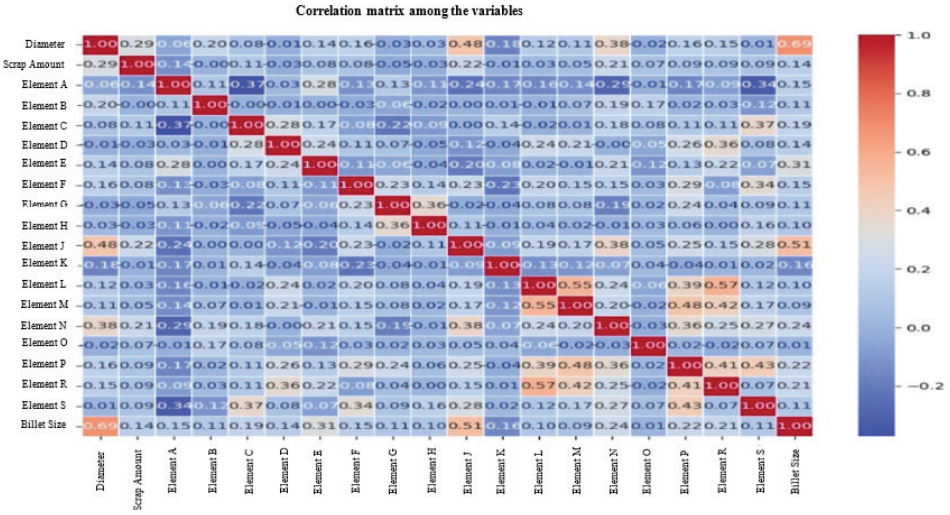


**Figure 2.** Correlation matrix among the variables.

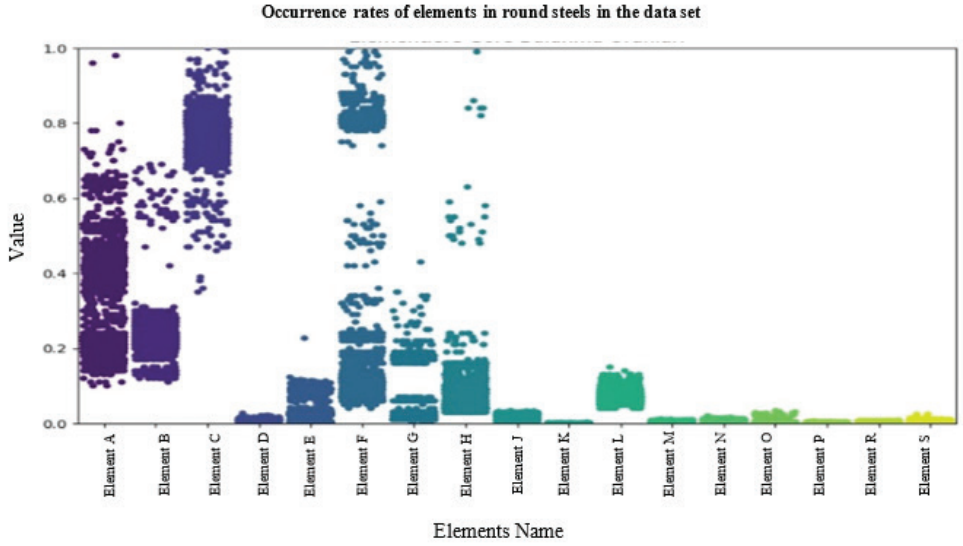


**Figure 3.** Occurrence rates of elements in round steels in the data set.

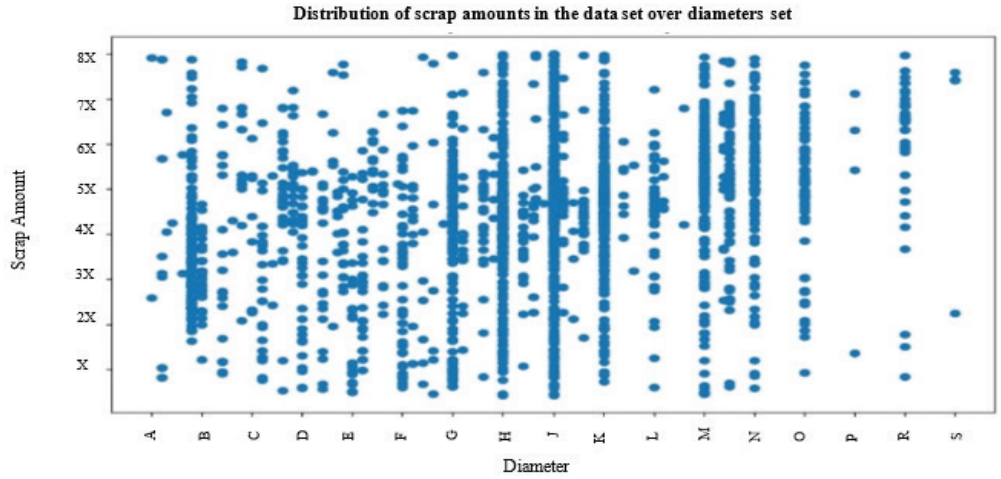


**Figure 4.** Distribution of scrap amounts in the data set over diameters.

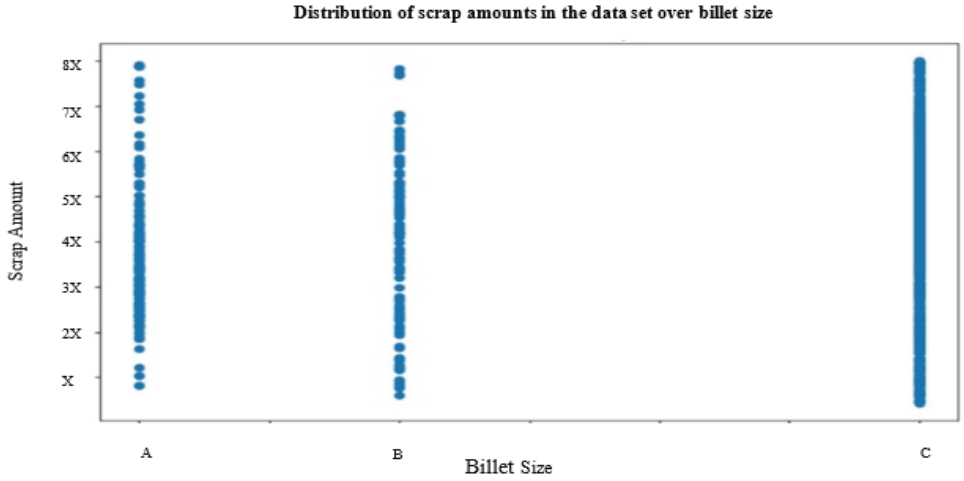


**Figure 5.** Distribution of scrap amounts in the data set over billet size.

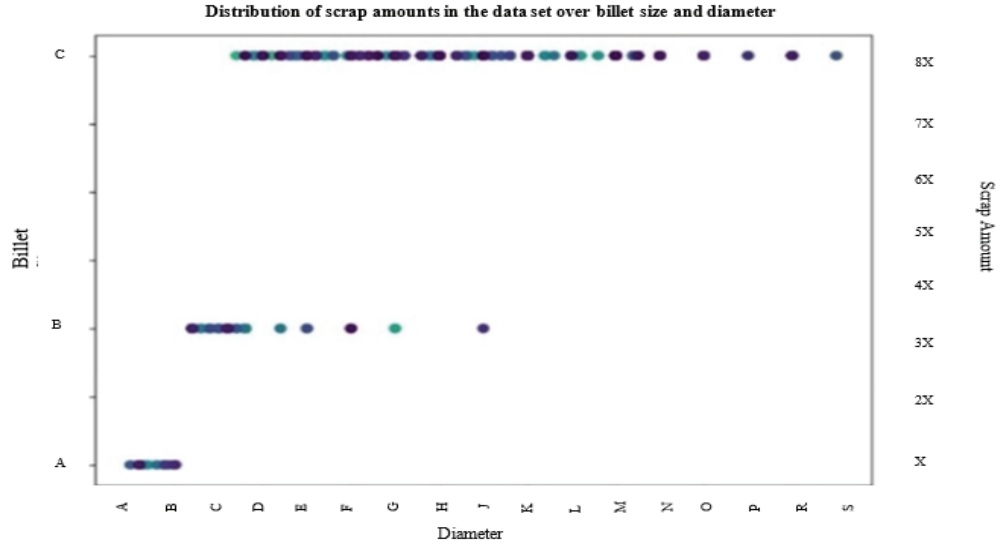


**Figure 6.** Distribution of scrap amounts in the data set over billet size and diameter.

The visualization of the data, to improve model fit and optimize error metrics, logarithmic data transformation has been applied across all algorithms and then the dataset was split into a 70% training set and a 30% test set for the application of machine learning algorithms. The dataset is divided into a training set and a test set; Training dataset, this subset of the data is used to train the machine learning model. It is utilized to fit the model's parameters and optimize its performance. Test dataset, this subset is reserved for evaluating the performance of the trained machine learning model. It is used to assess how well the model generalizes to new, unseen data [20].

In this study, a total of 4 regression algorithms and 3 classification algorithms were established and evaluated using model performance metrics. In the machine learning models established with regression algorithms, the prediction of scrap amount was carried out using the entire dataset. The models established with classification algorithms aimed to predict the scrap amount based on the diameter of the material and billet size. The algorithms used for the dataset are as follows.

**Multiple Linear Regression**

Multiple linear regression is an algorithm that establishes a linear relationship between multiple independent variables and a dependent variable, aiming to predict the outcome of future events [21]. This regression algorithm is designed to predict the value of the dependent variable using the values of multiple independent variables. The coefficients of the linear equation that best predict the value of the dependent variable are estimated [22]. Multiple Linear Regression allows for more accurate predictions when a single independent variable is insufficient. Figure 7 illustrates the working principle of the algorithm.

**Support Vector Machines**

Support vector machines (SVM) is an algorithm used for both regression and classification problems [23-24]. Support Vector Machines aim to partition the data points in the dataset into different classes within their respective regions using a hyperplane. It creates decision boundaries with hyperplanes to effectively classify data points. With the assistance of the hyperplane, data points are assigned to different classes [25]. Maximizing the margin between data points ensures finding the optimal decision boundary between classes. There can be multiple hyperplanes between classes [26]. Figure 8 illustrates the working principle of the algorithm.

**Random Forest Algorithm**

Random forest algorithm is used in both continuous and categorical large or small-sized datasets [27]. Random Forest algorithm is formed by combining decision trees, each trained separately on individual variables. Each decision tree is trained independently with a random subset of the dataset. Predictions are made for each sample by summing the predictions of each tree, and calculations are performed by averaging the predictions of the trees [28]. Random Forest algorithm can be used for both classification and regression problems. Figure 9 illustrates the working principle of the algorithm.

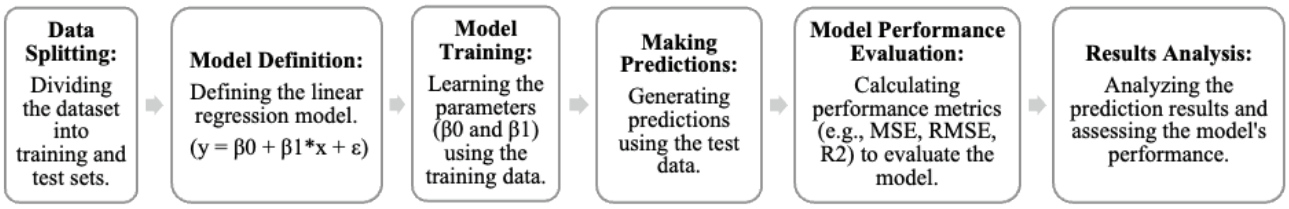


**Figure 7.** Multiple linear regression algorithm process flow chart.

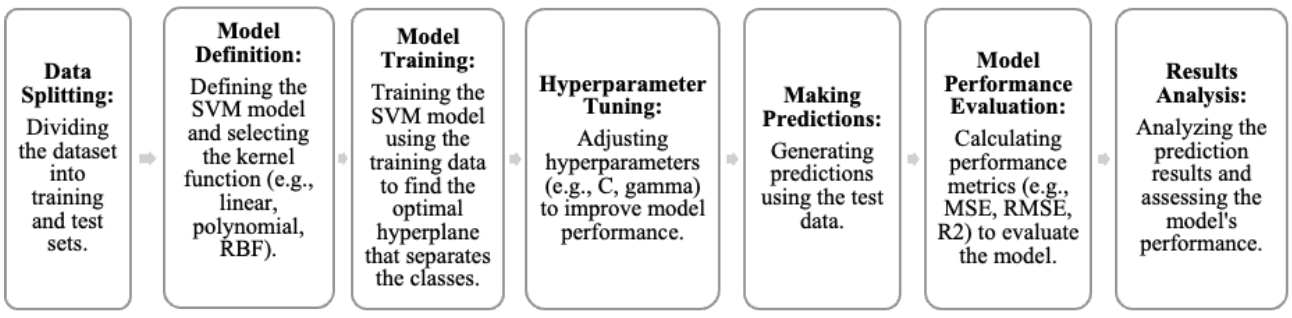


**Figure 8.** Support vector machines algorithm process flow chart.

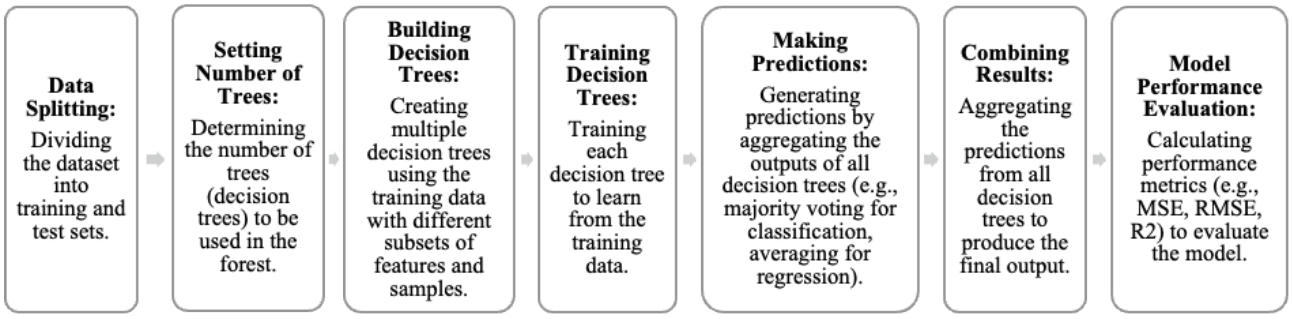


**Figure 9.** Random forest algorithm process flow chart.

**Ridge Regression**

Ridge regression is a regularized version of linear regression [29]. It is an algorithm used to address the problem of overfitting in regression models. It demonstrates resistance against the occurrence of overfitting [17]. Ridge Regression is an algorithm used for analyzing data where independent variables in a linear regression model exhibit high levels of fit and correlation. It is utilized to address the linearity-affected data. The aim is to minimize the difference between the observed data and the model's predictions, which are based on the training data. To mitigate overfitting and reduce bias, two main regularization techniques are employed, with the hyperparameter λ used for this purpose. This parameter controls the strength of the penalty applied to the model's coefficients. By tuning λ to its optimal value, Ridge Regression can demonstrate the best performance, particularly in addressing models prone to overfitting issues [30-31]. Figure 10 illustrates the working principle of the algorithm.

**K-Nearest Neighbours Algorithm**

K-nearest neighbours algorithm, commonly abbreviated as KNN, is used for both classification and regression problems [32]. The KNN algorithm performs prediction by

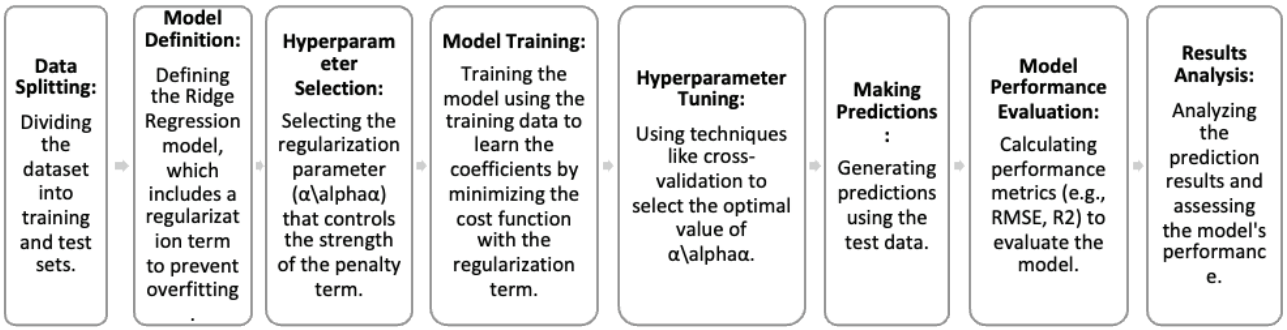


**Figure 10.** Ridge regression algorithm process flow chart.

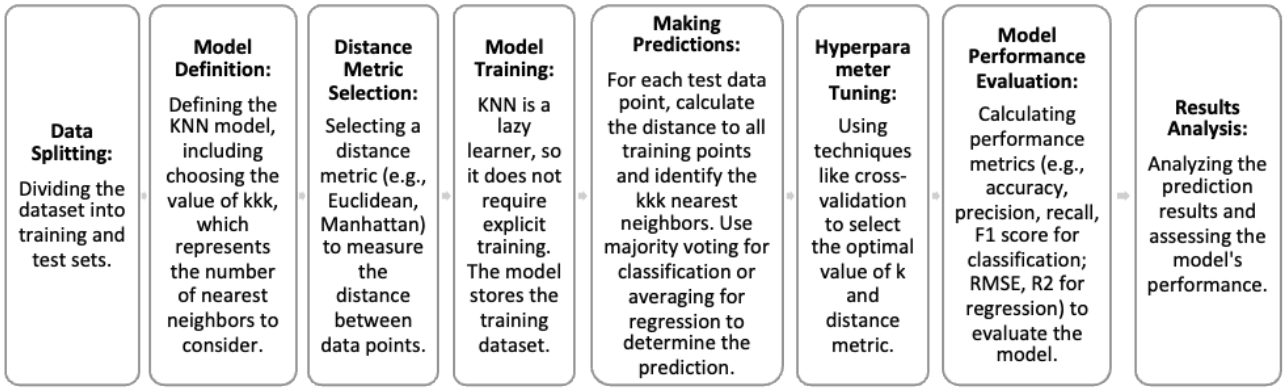


**Figure 11.** K-nearest neighbours algorithm process flow chart.

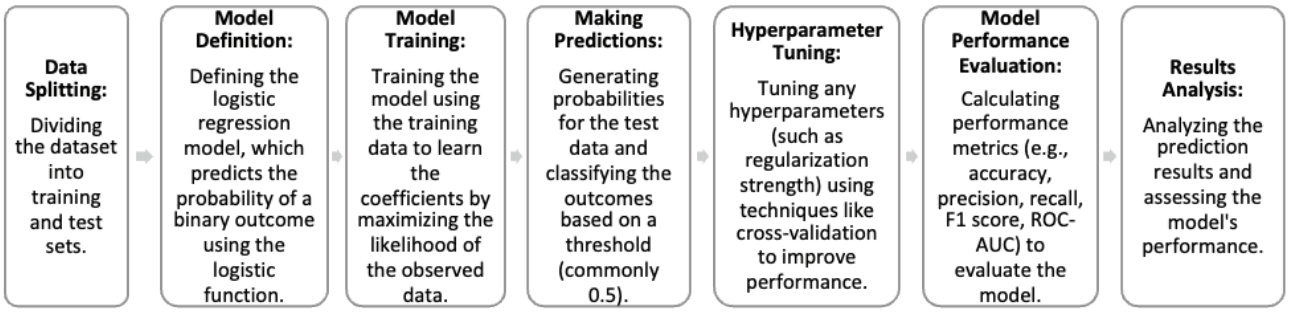


**Figure 12.** Logistic regression process flow chart.

determining which class the value to be predicted is concentrated in by identifying the nearest neighbours in the vector formed by the independent variables [33]. Figure 11 illustrates the working principle of the algorithm.

**Logistic Regression**

Logistic regression algorithm, despite its name containing the word "regression," is not a regression algorithm but a classification algorithm. It only deals with two possible outcomes probabilistically. It predicts the probability of belonging to a particular class and selects the highest probability. The predicted values lie between 0 and 1. the vector formed by the independent variables [34-35]. Figure 12 illustrates the working principle of the algorithm.

**Model Evaluation Metrics**

The performance of the models was evaluated using model performance metrics. In evaluating regression models, metrics such as R2, MAE, MSE, and RMSE are commonly used. R2 is a statistical measure of the agreement between actual data values and the predictions of the model [36]. MAE (Mean Absolute Error) calculates the mean of the absolute differences between the predicted values and the actual values [37]. MSE (Mean Squared Error) represents the average squared loss per sample in the entire dataset [38]. RMSE (Root Mean Squared Error) represents the square root of the mean squared error and standard deviation of predicted values errors [39]. The evaluation of classification models utilizes metrics such as accuracy, precision, recall, F1 score, confusion matrix ROC curve, and the associated Area Under the Curve (AUC). Accuracy represents the ratio of correctly predicted values in the dataset. Precision indicates how many of the predicted values are correctly classified. Recall represents the ratio of correctly predicted values to all true values in the dataset [40-41]. F1 score represents the harmonic mean of precision and recall. It can range from a minimum of 0 to a maximum of 1 [42]. In below (table 2) , confusion matrix is explained. Confusion matrix provides a concise summary of the model's performance by organizing predictions into categories such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents accurately classified positive examples, TN denotes accurately classified negative examples, FP indicates negative examples incorrectly classified as positive, and FN signifies positive examples incorrectly classified as negative. This matrix aids in assessing the accuracy and efficacy of the model's predictions by quantifying the agreement between predicted and actual values. [43-44].

**Table 2.** Confusion matrix

| | | Actual values | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted values** | Positive | TP | FP |
| | Negative | FN | TN |

Receiver operating characteristic (ROC) curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) on a graph. The true positive rate is calculated as the ratio of true positive predictions to the sum of true positives and false negative predictions, while the false positive rate is computed as the ratio of total false positive predictions to the sum of false positives and true negative predictions. Each point on the ROC curve corresponds to a different decision threshold for classification. The scale of the curve ranges from zero to one, where an ideal rate is represented by a TPR of one for positives and an FPR of zero for negatives. AUC (Area Under the Curve), represents the area under the ROC curve [45].

## RESULTS AND DISCUSSION

Following the analysis conducted on the dataset after removing outliers, the quality and diameters that yield the highest scrap amount were identified separately. The quality that yields the highest scrap amount based on the input quantity to the production area, the quality and diameters with the highest production output, and the element ratio within the material were analyzed. Upon this determination, a study will be carried out on the recorded items from relevant department. Additionally, the element that is more significant in affecting the scrap amount compared to other

**Table 3.** Regression models performance metrics comparison

| Method | R2 | MAE | MSE | RMSE |
|---|---|---|---|---|
| Multiple linear regression | 0.8996 | 123.414 | 35998 | 189.7314 |
| Support vector machine | 0.8829 | 0.0742 | 0.0235 | 0.1535 |
| Random forest algorithm | 0.9800 | 0.0452 | 0.6075 | 0.0866 |
| Ridge regression | 0.9966 | 32.920 | 2028.35 | 45.037 |

**Table 4.** Classification models performance metrics comparison

| Method | Accurancy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| Random forest algorithm | 0.6439 | 0.6134 | 0.5904 | 0.6017 | 0.6395 |
| K-nearest neighbour algorithm | 0.661 | 0.6718 | 0.7371 | 0.7029 | 0.6504 |
| Logistic regression | 0.6158 | 0.6157 | 0.7828 | 0.6893 | 0.6731 |

**Table 5.** Random forest algorithm confusion matrix

| | | Actual values | |
|---|---|---|---|
| | | Positive | Negative |
| **Predicted values** | Positive | 241 | 120 |
| | Negative | 109 | 173 |

**Table 6.** K-nearest neighbour algorithm confusion matrix

| | | Actual values | |
|---|---|---|---|
| | | Positive | Negative |
| **Predicted values** | Positive | 257 | 127 |
| | Negative | 93 | 166 |

**Table 7.** Logistic regression confusion matrix

| | | Actual Values | |
|---|---|---|---|
| | | Positive | Negative |
| **Predicted Values** | Positive | 274 | 171 |
| | Negative | 76 | 122 |

elements in the dataset was determined. The relationship between all variables was observed using a correlation matrix. In data analysis, in addition to the amount of scrap, it was also determined which quality and size consumed more energy, that analysis is for effect of the material price. Following the scrap amount data analysis, 7 machine learning algorithms were created, and the model performance metrics of all algorithms were evaluated.

A comparison of performance metrics, including R2, MAE, MSE, RMSE for regression problems, and accuracy, precision, recall, F1 score for classification problems, has been conducted and summarized in the table below (table 3 and table 4) and also expalained the confusion matrix all applied classification machine learning (table 5, table 6 and table 7)

It has been observed that among the regression models, random forest algorithm generally exhibits the best model performance across all metrics when evaluated. The best performance among classification models, when considering all metrics, it has been determined that the k-nearest neighbour algorithm.

After creating each machine learning algorithm and evaluating them using model performance metrics, an application design was developed using the Python programming language. This application design was created for both regression and classification algorithms. In the application design shown in Figure 12, for regression algorithms, parameters such as all elements, diameter, and billet size will be entered, and a prediction will be made using one of the regression algorithms selected from the dropdown menu. The results of all model performance metrics of the predictions made with the relevant data will be displayed in a popup window. The model performance metrics available in the popup window for regression algorithms are R2, MAE, MSE, and RMSE. For classification algorithms, parameters such as diameter and billet size will be entered, and a prediction will be made using one of the classification algorithms selected from the dropdown menu. The results of all model performance metrics of the predictions made with the relevant data will be displayed in a popup window. The model performance metrics available in the popup window for classification algorithms are accuracy, precision, recall, F1 score, and AUC. In addition to these metrics, the ROC curve for the relevant algorithm will also be plotted. The desging application screenshot is in below.

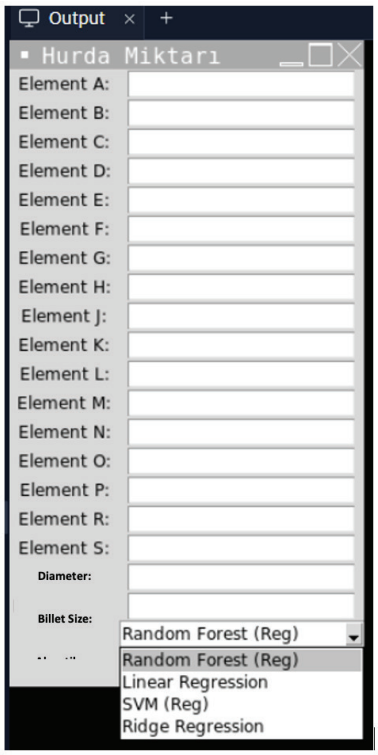


**Figure 13.** Design application screenshot.

## CONCLUSION

There are not many sources in the literature that utilize both data analytics and machine learning as two separate fields. In the evolving world, the increase in data and its interpretability has become crucial. With machine learning algorithms, meaningful relationships can be identified and predictions can be made more effectively. In this study, specifically focusing on the high-quality steel sector, prediction of scrap quantity in the selected rolling mill production area was conducted, and the significance of the presence ratios of elements on scrap quantity was investigated. The required amount of billets for the final product at the rolling mill production area, as well as the output quantity for the final product in terms of quality and dimensions, can be determined based on the input quantity to the production area. Predictions can be generated to ensure on-time delivery principles. This can prevent delivery delays to customers due to producing fewer products than forecasted. Detailed studies can also be conducted on the elements identified to have a significant impact on scrap quantity. At the end of the study, overall findings suggest that efforts can be made to reduce scrap quantities based on the identified factors. Alongside the creation and evaluation of these analyses and algorithms, an application has been developed. The application can be further enhanced and integrated into the relevant field or different departments, making it applicable across various areas. This study and similar research endeavors could explore different algorithms suitable for the dataset. Model performances can be compared, and evaluations can be expanded upon. The application was conducted only in the rolling mill production area of the high-quality steel production facility. Similar studies could be conducted in other production areas of the manufacturing plant. Furthermore, based on findings from other sectors, research on machine learning in the steel industry can be further expanded in the literature.

## AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

## DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

## CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## ETHICS

There are no ethical issues with the publication of this manuscript.

## STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

## REFERENCES

[1] Worldsteel. Total production of crude steel/world total 2023. Worldsteel. Available at: https://worldsteel.org/steel-topics/statistics/annual-production-steel-data/?ind=P1_crude_steel_total_pub/CHN/IND Accessed on March 12, 2024.

[2] Medya M. Vasıflı çelikler nedir? Demir Çelik Store. Available at: https://demircelikstore.com/urun/vasifli-celikler-nedir Accessed on March 12, 2024.

[3] Çelik UD. Vasıflı çelik. Uslu Demir Çelik. Available at: https://www.usludemircelik.com/vasiflicelik.html Accessed on March 12, 2024.

[4] Kang Z, Catal C, Tekinerdogan B. Machine learning applications in production lines: A systematic literature review. Comput Ind Eng 2020;149:106773.

[5] Koruyan K, Ekeryılmaz A. Makine öğrenmesi ile müşteri şikayetlerinin sınıflandırılması. Acad J Inform Technol 2022;13:50.

[6] Mirasçı S, Aksoy A. Data mining and machine learning applications in steel materials purchasing. J Eng Sci Design 2023;11:1174–1189.
[7] Qian X, Suvarna M, Li J, Zhu X, Cai J, Wang X. Online prediction of mechanical properties of hot rolled steel plate using machine learning. Mater Des 2021;197:109201.
[8] Sandhya N, Sowmya V, Bandaru CR, Babu GR. Prediction of mechanical properties of steel using data science techniques. Int J Recent Technol Eng 2019;8:235–41.
[9] Raju SMTU, Sarker A, Das A, Islam M, Al-Rakhami SM, Al-Amri AM, Mohiuddin T, Albogamy FR. An approach for demand forecasting in steel industries using ensemble learning. Complexity 2022;2022:1–19.
[10] Chahbi I, Rabah NB, Tekaya IB. Towards an efficient and interpretable machine learning approach for energy prediction in industrial buildings: A case study in the steel industry. IEEE 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2022, pp. 1-8.
[11] Choi S, Seo B, Lee E. Machine learning-based tap temperature prediction and control for optimized power consumption in stainless electric arc furnaces (EAF) of steel plants. Sustainability 2023;15:1–31.
[12] Shiraiwa T, Miyazawa Y, Enoki M. Prediction of fatigue strength in steels by linear regression and neural network. Mater Trans 2018;60:189–198.
[13] Patel SV, Jokhakar VN. A random forest based machine learning approach for mild steel defect diagnosis. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).
[14] Ecemiş O, Irmak S. Comparison of data mining methods in stainless steel sector sales forecasting. Kilis Univ Soc Sci 2018;8:148–169. [Turkish]
[15] Byjus. Data sets. Byjus. Available at: https://byjus.com/maths/data-sets Accessed on March 12, 2024.
[16] Crabtree M, Nehme A. What is data analysis? An expert guide with examples. Datacamp. Available at: https://www.datacamp.com/blog/what-is-data-analysis-expert- Accessed on October 15, 2023.
[17] Dikengül H. Türkiye'deki son 100 yıldaki depremlerin makine öğrenmesiyle analizi (Yüksek Lisans Tezi). Ankara: Ufuk Üniversitesi; 2023.
[18] IBM. What is machine learning (ML)? IBM. Available at: https://www.ibm.com/topics/machine-learning Accessed on March 12, 2024
[19] McKinney W. Python for data analysis. O'Reilly; 2012.
[20] Brownlee J. Train-test split for evaluating machine learning algorithms. Machine Learning Mastery. Available at: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms Accessed on March 12, 2024.
[21] Kanade V. What is linear regression? Types, equation, examples, and best practices for 2022. Spicework. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression Accessed on November 24, 2023.
[22] IBM. What is linear regression? IBM. Available at: https://www.ibm.com/topics/linear-regression Accessed on October 16, 2023.
[23] Mathworks. Understanding support vector machine regression. Mathworks. Available at: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html Accessed on February 6, 2024.
[24] Scikit-Learn. Support vector machines. Scikit-Learn. Available at: https://scikit-learn.org/stable/modules/svm.html Accessed on February 6, 2024.
[25] Gandhi R. Support vector machine — Introduction to machine learning algorithms. Medium. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 Accessed on November 14, 2023.
[26] IBM. What are support vector machines (SVMs)? IBM. Available at: https://www.ibm.com/topics/support-vector-machine Accessed on February 6, 2024.
[27] Watts JD, Powell SL, Lawrence RL, Hilker TL. Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. Remote Sens Environ 2011;115:66–75.
[28] Breiman L. Random Forests. Dordrecht, Netherlands: Kluwer Academic Publishers; 2001. p. 5–32.
[29] Eren M. İstanbul metro hatları için makine öğrenmesi ile yolcu sayılarının tahminlenmesi (Yüksek Lisans Tezi). Eskişehir: Eskişehir Osmangazi Üniversitesi; 2022.
[30] Lago B. Mastering ridge regression: A key to taming data complexity. Medium. Available at: https://medium.com/@bernardolago/mastering-ridge-regression-a-key-to-taming-data-complexity-98b67d343087 Accessed on December 23, 2023.
[31] IBM. What is ridge regression? IBM. Available at: https://www.ibm.com/topics/ridge-regression#:~:text=Ridge%20regression%20is%20a%20statistical,regularization%20for%20linear%20regression%20models Accessed on March 12, 2024.
[32] IBM. What is the KNN algorithm? IBM. Available at: https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today Accessed on March 12, 2024.
[33] Zilyas D. Makine öğrenmesi yöntemi ile eğitim başarısının tahmini modeli (Yüksek Lisans Tezi). İstanbul: İstanbul Beykent Üniversitesi; 2023.
[34] Swaminathan S. Logistic regression — Detailed overview. Medium. Available at: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc Accessed on December 23, 2023.

[35] Thomas WE, Manz DO. Exploratory study. In: Batarseh FA, Yang R, (editors). Research Methods for Cyber Security. Syngress; 2017. p. 95–130.

[36] J F. R-squared: Definition, calculation formula, uses, and limitations. Investopedia. Available at: https://www.investopedia.com/terms/r/r-squared.asp Accessed on March 12, 2024.

[37] Schneider P, Xhafa F. Mean absolute error (MAE). In: Xhafa F, editor. Anomaly detection and complex event processing over IoT data streams. Cambridge, Massachusetts: Academic Press; 2022. p. 49–66.

[38] Kozyrkov C. How to use the MSE in data science. Medium. Available at: https://medium.com/@kozyrkov/how-to-use-the-mse-in-data-science-bd350154a9d Accessed on March 12, 2024.

[39] Statisticshowto. RMSE: Root mean square error. Statisticshowto. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error Accessed on March 12, 2024.

[40] Evindentlyai. Accuracy vs. precision vs. recall in machine learning: What's the difference? Evindentlyai. Available at: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall#:~:text=Recall%20is%20a%20metric%20that,the%20number%20of%20positive%20instances Accessed on March 12, 2024.

[41] Koehrsen W. Precision and recall: How to evaluate your classification model. Builtin. Available at: https://builtin.com/data-science/precision-and-recall Accessed on March 12, 2024.

[42] Sharma N. Understanding and applying F1 score: AI evaluation essentials with hands-on coding example. Arize. Available at: https://arize.com/blog-course/f1-score/#:~:text=F1%20score%20is%20a%20measure,can%20be%20modified%20into%20F0 Accessed on March 12, 2024.

[43] Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. In: Batarseh FA, Yang R, editors. Data Democracy. Cambridge, Massachusetts: Academic Press; 2020. p. 83–106. Accessed on March 12 2024.

[44] Encord. Confusion matrix. Encord. Available at: https://encord.com/glossary/confusion-matrix/ Accessed on March 12, 2024.

[45] Coursera Staff. What is ROC curve in machine learning? Coursera. Available at: https://www.coursera.org/articles/what-is-roc-curve Accessed on March 12, 2024.