Physicochem. Probl. Miner. Process., 59(5), 2023, 169823

http://www.journalssystem.com/ppmp

Physicochemical Problems of Mineral Processing

ISSN 1643-1049

Developing a data-driven soft sensor to predict silicate impurity in iron ore flotation concentrate

Yusuf Enes Pural

Istanbul Technical University, Faculty of Mines, Mineral Processing Engineering Department, Maslak, Istanbul, Turkey

Corresponding author: pural@itu.edu.tr (Yusuf Enes Pural)

Abstract: Soft sensors are mathematical models that estimate the value of a process variable that is difficult or expensive to measure directly. They can be based on first principle models, data-based models, or a combination of both. These models are increasingly used in mineral processing to estimate and optimize important performance parameters such as mill load, mineral grades, and particle size. This study investigates the development of a data-driven soft sensor to predict the silicate content in iron ore reverse flotation concentrate, a crucial indicator of plant performance. The proposed soft sensor model employs a dataset obtained from Kaggle, which includes measurements of iron and silicate content in the feed to the plant, reagent dosages, weight and pH of pulp, as well as the amount of air and froth levels in the flotation units. To reduce the dimensionality of the dataset, Principal Component Analysis, an unsupervised machine learning method, was applied. The soft sensor model was developed using three machine learning algorithms, namely, Ridge Regression, Multi-Layer Perceptron, and Random Forest. The Random Forest model, created with non-reduced data, demonstrated superior performance, with an R-squared value of 96.5% and a mean absolute error of 0.089. The results suggest that the proposed soft sensor model can accurately predict the silicate content in the iron ore flotation concentrate using machine learning algorithms. Moreover, the study highlights the importance of selecting appropriate algorithms for soft sensor developments in mineral processing plants.

Keywords: soft sensor, machine learning, random forest, multi-layer perceptron, flotation, grade estimation

1. Introduction

In mineral processing, one of the key challenges is to extract meaningful patterns from raw data. Machine learning techniques have been widely used to develop data-driven models that can describe the system based on historical data using different algorithms. These models, often referred to as soft sensors in the process industry, allow predicting variables that cannot be measured directly but are key performance indicators, using process variables that are easy to measure. In the context of flotation, soft sensors have been used for various purposes such as mill filling determination, particle size and mineral grade estimation, fault detection, etc. An extensive review by McCoy & Auret (2019) highlights the importance of soft sensors in the mineral processing industry, particularly in flotation.

Several recent studies have utilized machine learning techniques to develop soft sensors for flotation. Popli et al. (2018) conducted experiments to examine the effect of particle size and KEX dosages in each circuit using factorial design in Lead-Zinc flotation. Li, Gui, and Zhu (2019) developed a hybrid CNN-Support Vector Machine model using froth images to check the suitability of reagent dosages. Jian, Lihui, and Xie (2020) developed a CNN model using froth pictures in antimony flotation to estimate grade. Wen et al. (2021) used convolutional neural networks to estimate the ash content of the clean coal by taking photographs from an industrial-sized cell in coal flotation. Zhang and Gao (2021) used 13 different deep learning algorithms to determine the iron content in the flotation residue. Montanares et al. (2021) used three different machine learning algorithms to determine the silicate grade, where the LSTM model showed the best performance. Gomez-Flores et al. (2022) utilized k-nearest neighbors,

DOI: 10.37190/ppmp/169823

decision trees, and random forest models to estimate grade and recovery in flotation based on 18 variables and 330 measurements from 19 articles. Qi, Za, and Ga (2022) determined the iron content in the waste by taking images of the froth in an iron ore reverse flotation and modeled the data obtained by image processing with a multilayer perceptron. Zhang and Gao (2022) determined the reagent dosages based on the estimation made in the previous article, using an extreme learning machine algorithm including some variables in flotation. Ren et al. (2022) developed the LSTM model to predict the gangue mineral content in the flotation concentrate. Finally, Zhao et al. (2022) classified coals as refined or tailing using four different machine learning algorithms based on more than 250,000 particle MLA data, where the random forest model showed the best performance with an AUC score of 0.97.

In this study, a soft sensor was developed using data-driven methods to accurately predict the silicate content in iron ore flotation concentrate, highlighting the importance of appropriate algorithm selection for modelling of soft sensor in mineral processing plants.

2. Dataset and algorithms

2.1. Dataset and exploratory data analysis

The data of the flotation process used in the study were taken from the Kaggle platform (Oliveira, 2017). In order to enrich the iron, reverse flotation is performed and while the iron is depressed with starch, the silicate is floated using amine type collector in an alkaline media.

The dataset contains 21 inputs, 1 output variable and more than 700,000 measurements. Brief descriptive statistics of inputs and output are given in Table 2. The iron and silicate content in the feed is determined by an average of 14 hours of accumulated sample. Other inputs are measured by sensors every 20 seconds. The silicate content in the concentrate, which is the response variable, is analysed hourly.

Table 2. Summary of descriptive statistics

Features	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
% Iron in Feed	736282	56.3	5.2	42.7	52.7	56.1	59.7	65.8
% Silica in Feed	736282	14.6	6.8	1.3	8.9	13.9	19.6	33.4
Starch Flow, m^3/h	736282	2869.6	1216.0	0.0	2075.1	3020.2	3728.9	6300.2
Amina Flow, m^3/h	736282	488.2	91.3	241.7	431.8	504.4	553.3	739.5
Ore Pulp Flow, t/h	736282	397.6	9.7	376.2	394.2	399.2	403.0	418.6
Ore Pulp pH	736282	9.8	0.4	8.8	9.5	9.8	10.0	10.8
Ore Pulp Density, t/m^3	736282	1.7	0.1	1.5	1.6	1.7	1.7	1.9
FC 01 Air Flow, Nm^3/h	736282	280.1	29.6	175.5	250.3	299.3	300.1	373.9
FC 02 Air Flow	736282	277.1	30.2	175.2	250.4	296.2	300.7	376.0
FC 03 Air Flow	736282	281.1	28.6	176.5	250.8	298.7	300.4	364.3
FC 04 Air Flow	736282	299.4	2.6	292.2	298.3	299.8	300.6	305.9
FC 05 Air Flow	736282	299.9	3.6	286.3	298.1	299.9	301.8	310.3
FC 06 Air Flow	736282	292.1	30.2	189.9	260.3	299.5	303.1	370.9
FC 07 Air Flow	736282	290.7	28.7	186.0	256.0	299.0	301.9	371.6
FC 01 Froth Level, mm	736282	520.2	131.1	149.2	416.9	491.7	594.1	862.3
FC 02 Froth Level	736282	522.6	128.2	210.8	441.8	495.9	595.3	828.9
FC 03 Froth Level	736282	531.3	150.9	126.3	411.3	494.2	601.3	886.8
FC 04 Froth Level	736282	420.2	91.8	162.2	356.6	411.8	485.3	680.4
FC 05 Froth Level	736282	425.1	84.5	167.0	357.6	408.7	484.0	675.6
FC 06 Froth Level	736282	429.9	89.9	155.8	358.4	424.4	492.8	698.9
FC 07 Froth Level	736282	420.9	84.9	175.3	356.7	410.9	476.1	659.9
% Silica in Concentrate	736282	2.3	1.1	0.6	1.4	2.0	3.0	5.5

Distributions of all values for features are given in Fig. 1 as box plot for better evaluation. It can be seen from the Fig. that the scales of the features are different from each other. For example, pulp density ranges from 1.5 to 1.9, while amine dosage ranges from 240 to 740. In situations where the features are characterized by varying scales, many machine learning algorithms may fail to produce accurate results. Hence, to address this issue effectively, scaling the data prior to the modeling phase becomes an indispensable necessity (Ahsan et al., 2021). By standardizing the features to a common scale, we can ensure that the machine learning models can perform optimally and provide more reliable and meaningful insights from the data. Also, some features have a large number of outliers. These outliers can signify that the feature of interest possesses extreme values or might have arisen due to errors originating from sensors or analytical procedures. Consequently, these specific data points have the potential to bias the model training process and consequently lead to less precise predictions (Fernandez et al., 2022). Addressing outliers through appropriate data preprocessing techniques becomes crucial to ensure the robustness and accuracy of the predictive model. Moreover, there is a high correlation between some of the features that can be seen in Fig. 2. The initial salient observation in this context is the presence of a negative correlation between the iron content and silica content within the feed. Furthermore, an observed positive correlation emerges between the air velocities and froth levels within the flotation cells. Such a multicollinearity, a phenomenon in which independent variables in a regression model are highly correlated, can have a significant impact on the estimation of the coefficients. When a multicollinearity is present, it causes the standard errors to increase, making some or all of the coefficients for the independent variables appear statistically insignificant, even if they should be considered significant contributions to the model (Daoud, 2017). In such cases, the use of some algorithms is not appropriate or it is necessary to apply data preprocessing before modelling.

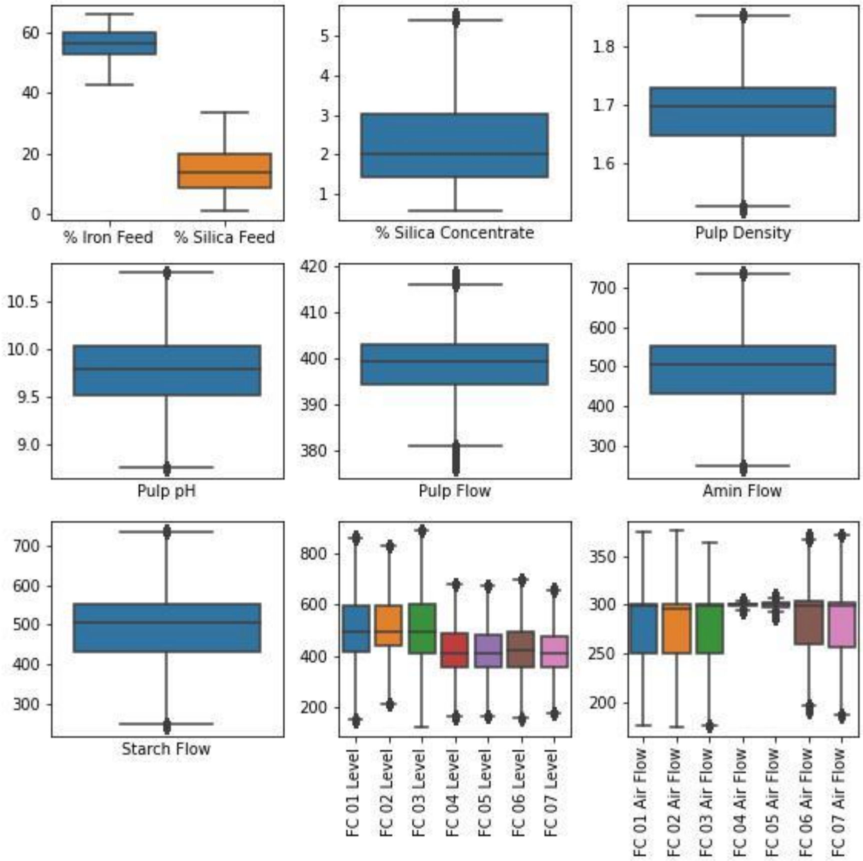


Fig. 1. Boxplots of all features

2.2. Data Preprocessing

Measurements to be used to determine model performance should not be in the dataset to be used in model training. That is, the test of a model performance should always be done exclusively on unseen data. For this purpose, the original dataset was randomly allocated as 75% training and 25% testing. In addition, 5-fold cross validation was used in the training data for model hyperparameters tuning. The data splitting method is given in Fig. 3.

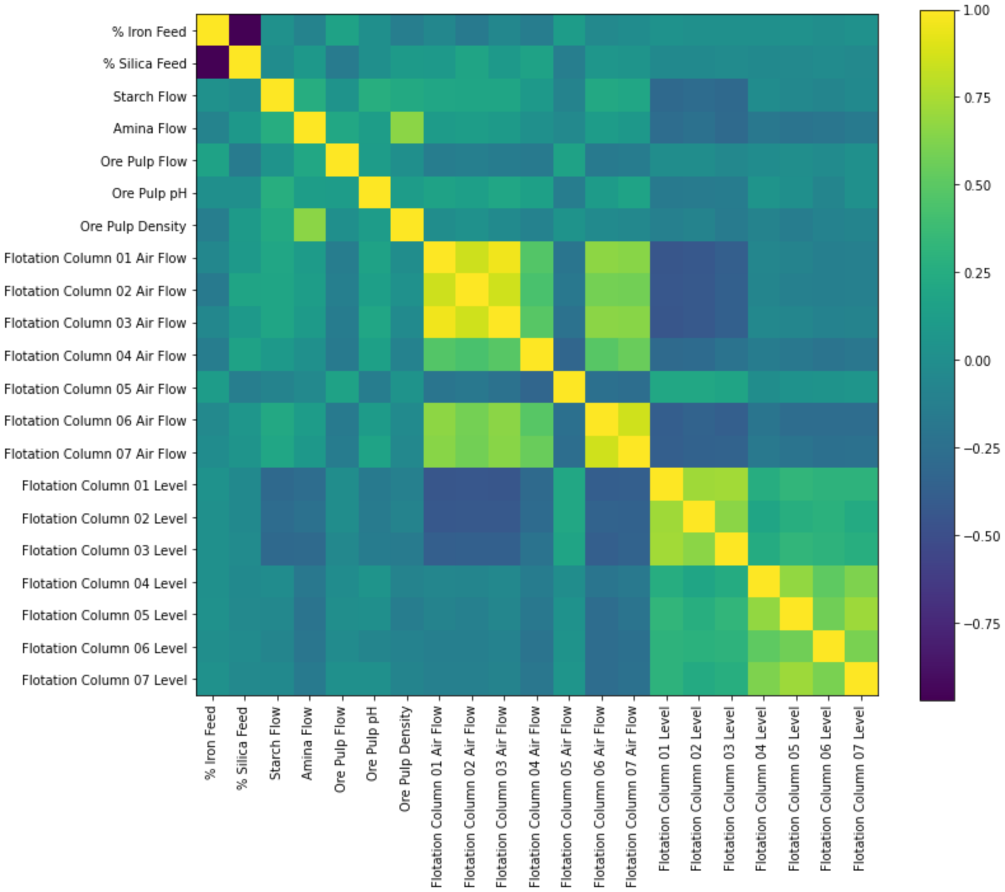
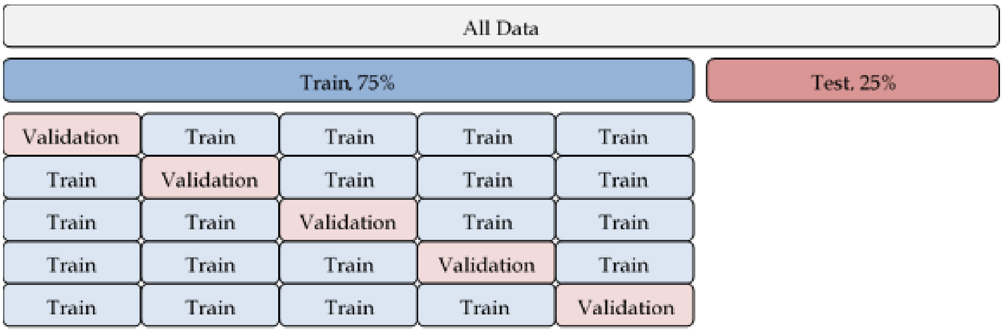


Fig. 2. Input features correlation matrix



All Data				
Train, 75%				Test, 25%
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train
Train	Train	Validation	Train	Train
Train	Train	Train	Validation	Train
Train	Train	Train	Train	Validation

Fig. 3. Data splitting as train, validation and test

Before using Multi-Layer Perceptron model (MLP), Principal Component Analysis (PCA) was applied to the data. PCA is one of the unsupervised machine learning methods. Since the input variables can be represented with fewer features with PCA, estimation errors due to both noisy data and correlation between features can be minimized. Also, machine learning algorithms converge faster with reduced inputs.

PCA works sensitive to the scale of the features. Therefore, data standardization (z-score normalization) was performed using Equation 1 for feature scaling, where z is a standard score, x_i is the original value, μ is mean of the sample, and s is the standard deviation of a sample.

$$z = \frac{x_i - \mu}{s} \tag{1}$$

In Fig. 4 it appears that 90% of the variability in 21 inputs can actually be explained by 11 features. Thus, instead of 21 columns of data, 11 columns of data will be fed with 10% information loss to train the multi-layer perceptron model, which is one of the artificial neural network models that require high processing power.

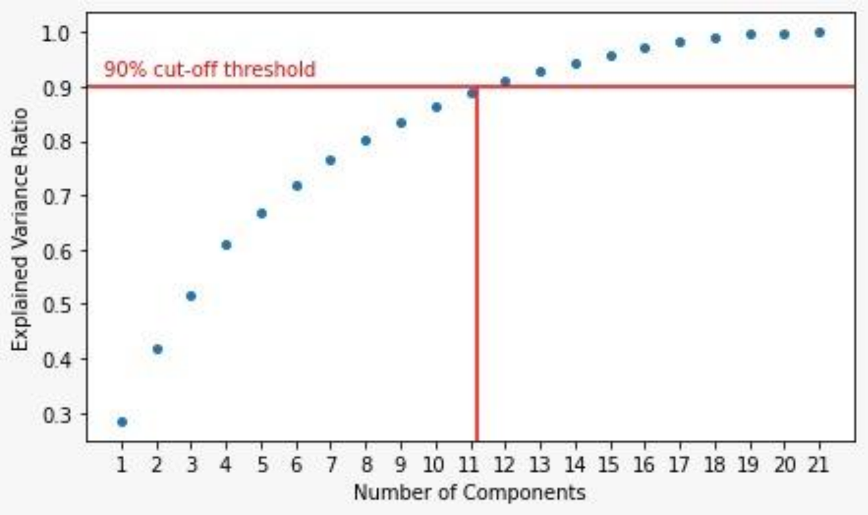


Fig. 4. Cumulative variance explanations by principal components

2.3. Modelling

In mineral processing, especially flotation, the process is often very complex and there is no linear model to describe the process. For this reason, MLP and the Random Forests (RF) methods, which have high predictive power and are more successful in explaining complex systems, have been used as well as ridge regression.

2.3.1. Ridge regression

Linear regression is a statistical method for modeling the relationship between a dependent variable and independent variable. The model can be expressed in equation form as:

$$Y = X\hat{\beta} + \epsilon$$

where Y is the dependent variable, X is the independent variable, $\hat{\beta}$ is the coefficient, and ϵ is the error term. The objective of simple linear regression is to estimate the values of $\hat{\beta}$ that minimize the sum of squared errors between the predicted and actual values of Y. The final equation for the estimated coefficients is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Ridge regression, on the other hand, is a modification of linear regression that is used when the independent variables are highly correlated with each other. In this case, the estimates of the coefficients in linear regression can be unstable and may have large variances. Ridge regression adds a penalty term to the regression equation that shrinks the estimated coefficients towards zero, which helps to reduce the variance of the estimates. The penalty term is controlled by a tuning parameter, λ, which determines the strength of the penalty. The final equation for the estimated coefficients is:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

where I is the identity matrix. The presence of high correlation among some independent variables as shown in Fig. 2, necessitates the use of ridge regression instead of linear regression in this study. To find the optimal ridge lambda coefficient, a set of alpha values was chosen for experimentation. These alpha values were specifically generated using a function that generated six logarithmically spaced values within the range from 0.001 to 1000. After evaluating the model's predictive accuracy at each alpha value, the optimal ridge lambda coefficient was determined as 63,1.

2.3.2. Multi-layer perceptron

MLP is a fully connected feedforward artificial neural network and consists of three types of layers as shown in Fig. 5. Input layer contains the neurons in which features are represented. The hidden layers are located between the input and output layers. Neurons here take in a set of weighted inputs and produce an output through an activation function. The output layer receives the values from the last hidden layer and transforms them into output values.

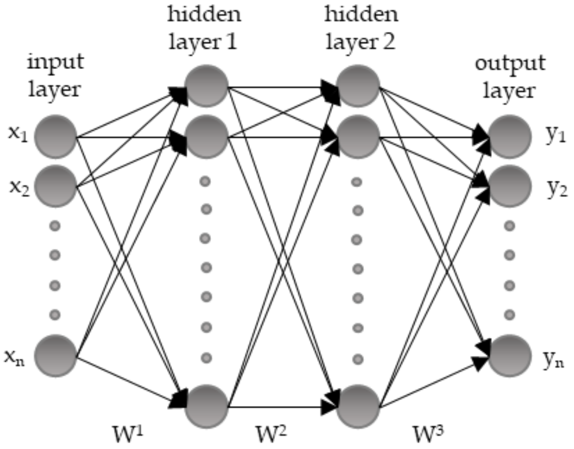


Fig. 5. Structure of Multi-Layer Perceptron

The performance of the MLP is highly dependent on number of hidden layers and number of neurons in each layer. In addition, there are many hyperparameters that needs to be optimized such as activation function, learning rate and solver method etc. However, no hyperparameter optimization has been done to reduce the computation time.

In this study, 3 hidden layers, each containing 500 neurons, were used to train the model, based on past experience. For the optimization of the weights, Adam Optimizer, which is a gradient descent-based algorithm that gives successful results on large data sets, was used. Rectifier Linear Unit (ReLU), whose graph is given in Fig. 6, was chosen as the activation function.

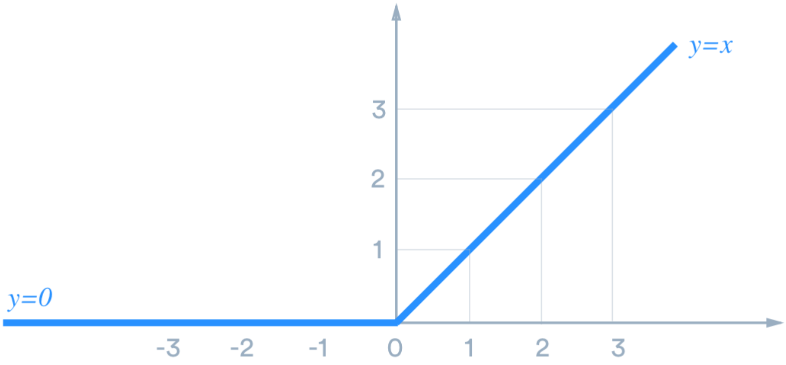


Fig. 6. Structure of Multi-Layer Perceptron

2.3.3. Random Forest

Random Forest is an ensemble learning method used in both classification and regression applications. In this method, many sub-datasets are obtained using bootstrap sampling from original dataset. Unlike some other ensemble learning methods, features are also randomly selected while creating the sub-dataset. Each sub-dataset is then used to train the decision tree model. Finally, the final estimate is made by averaging the results from all decision trees. The visualization of random forest is given in Fig. 7.

The RF can naturally deal with features at different scales, outliers, and collinearity. For this reason, no data preprocessing steps were applied before training the model.

There are many hyperparameters in the random forest model. Most important ones are as follow with their tested interval and explained in Fig. 8 on a single tree.

- the number of trees in the forest (100, 200, 300, 400),
- the maximum depth of the tree (300, 400, 500),
- the minimum number of samples required to split an internal node (2, 5, 10) and

the minimum number of samples required to be at a leaf node (2, 5, 10).

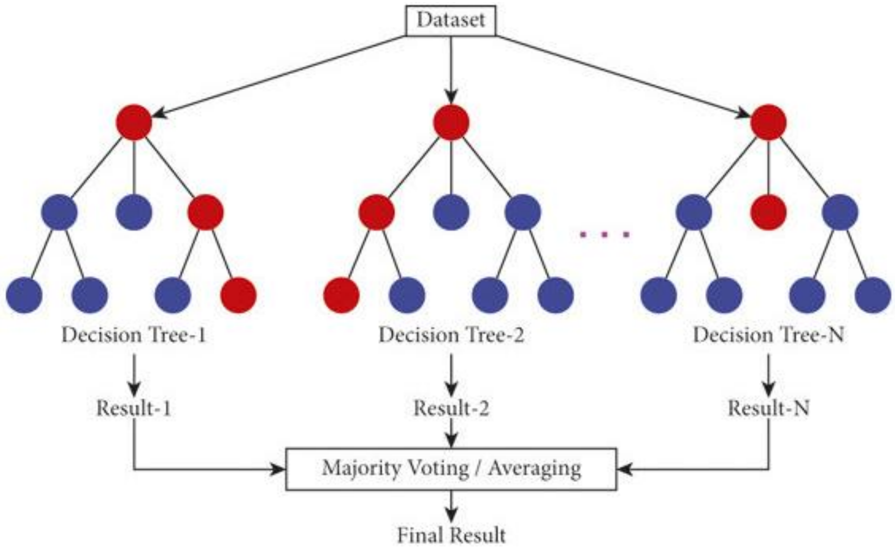


Fig. 7. Illustriation of random forest

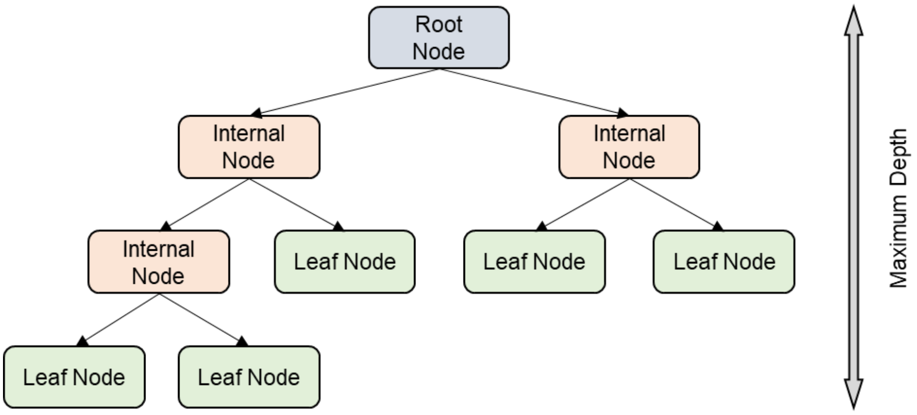


Fig. 8. Structure of a single tree

In this study, grid search method was used together with the previously mentioned 5-fold cross validation for the optimization of random forest model hyperparameters. The best combination of hyperparameters was found as 300 for number of trees, 500 for maximum depth, 2 for both minimum sample for split and minimum sample in leaf.

As stated before, in RF, features are randomly selected at each node. In this way, by looking at how much the tree nodes using an attribute reduce the impurity on average, the importance of that attribute can be measured. In Fig. 9, the importance levels of the features are given. From the graph, it can be seen that the amount of silica impurity in the concentrate is highly dependent on the iron and silicate content in the fed ore. Additionally, it is understood that flotation pH and amine dosage are among the most important factors. These findings highlight the critical role of these variables in the concentration process and emphasize their significance in optimizing the flotation performance for effective silica impurity removal.

3. Results and discussion

This study covers estimation of residual content in flotation concentrate using some machine learning algorithms. For this purpose, more than 700 thousand data of 21 properties obtained from the industrial flotation plant, shared publicly, were used. After applying the necessary data preprocessing steps, Ridge Regression, Random Forest and Multi-Layer Perceptron models were trained with 75% of the original data set. The other 25% of the data, which the models had not seen before, was used to test the prediction performance of the models. R^2 and mean absolute error (MAE) were selected to evaluate model performance. The results are presented in Table 3.

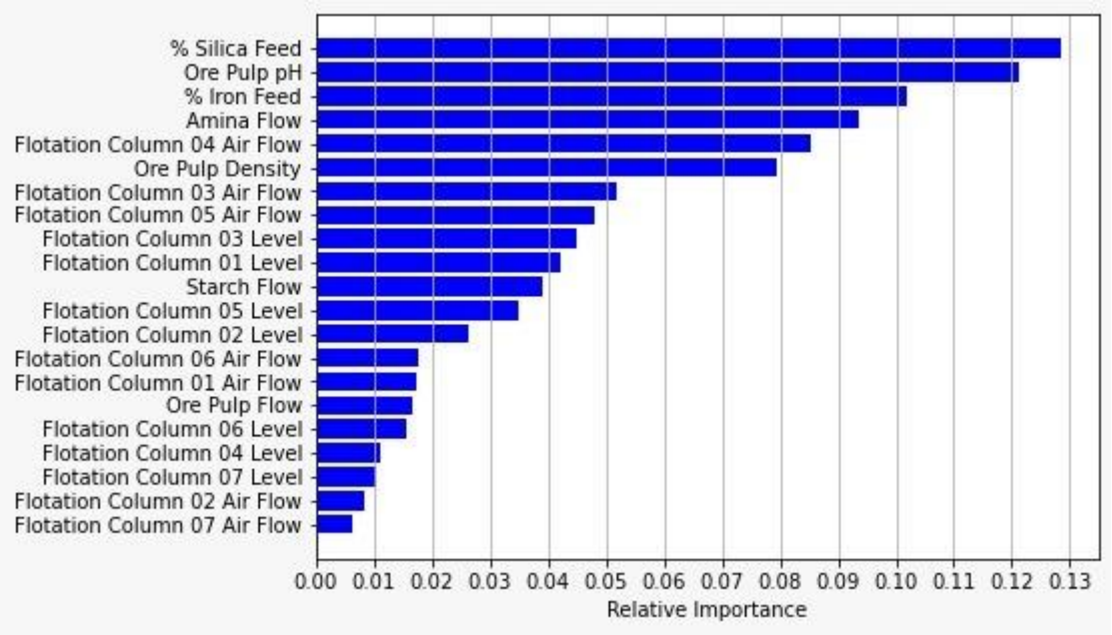


Fig. 9. Importance level of features

Table 3. Performance of tested models

Model	R^2	MAE
Ridge Regression	0.153	1.213
Random Forest	0.965	0.089
Multi-Layer Perceptron	0.886	0.243

The superior model is RF with higher R^2 and lower MAE as can be seen from Table 3. This may be because the hyperparameters of the RF model are fine-tuned. Of course, model success can be increased by fine-tuning the MLP hyperparameters. However, since RF is successful enough and no data preprocessing step is required for RF, it can be said that no further processing is required.

By using the model, the operating conditions can be optimized to meet the constraints defined in the next stages. For example, depending on the varying iron and silicate content of the feed, optimum levels of inputs such as collector dosage or pulp pH can be determined. The study by Fan et al. (2020) brought attention to the diverse forms of dodecylamine (DDA) found in aqueous solutions across different pH levels, particularly highlighting its efficacy in the reverse flotation of iron ore with amines at the optimal pH of 10.5. At this pH, the concentration of the RNH2 ·RNH3+ ion-molecule complex attains its highest level. Within our dataset, the pH values exhibited a range from 8.8 to 10.8, with an average of 9.8. Notably, this dataset encompassed a significant variation in the silica content of the concentrate, ranging from as low as 0.6% to as high as 5.5%. This substantial fluctuation underscores the substantial influence of pH as one of the most pivotal factors affecting the flotation process, as clearly demonstrated in Fig. 9.

4. Conclusion

This study underscores the significance of soft sensors in mineral processing plants as powerful tools for identifying key performance indicators that are difficult or costly to measure directly. Specifically, it was successfully developed a soft sensor, utilizing the Random Forest method, to estimate grade in the flotation plant with an exceptional R-squared value of 96.5%. The insights gained from analyzing the importance levels of input variables have provided a deeper understanding of the underlying processes, enabling better control and optimization of the flotation process. Soft sensors offer a promising avenue for similar applications across all mineral processing operations, providing real-time estimations and invaluable data-driven insights. Through accurate data measurement and the appropriate implementation of soft sensor algorithms, mineral processing plants can enhance their operational efficiency, reduce costs, and improve overall performance. As the mining industry continues to embrace the potential of soft sensors, these intelligent tools are set to revolutionize mineral processing practices, bringing about significant advancements in process control, optimization, and decision-making.

Acknowledgements

In this study, National Center for High Performance Computing (UHEM) machines were used to develop machine learning models with grant number of 011592022.

References

ABUSNİNA, A., 2014. *Gaussian Process Adaptive Soft Sensors and their Applications in Inferential Control Systems.* EngD. Thesis.

AHSAN, M. M., MAHMUD, M. A. P., SAHA, P. K., GUPTA, K. D., & SİDDİQUE, Z., 2021. *Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance*. Technologies. 9(3), 5–9.

DAOUD, JAMAL I., 2019. *Multicollinearity and Regression Analysis.* Journal of Physics: Conference Series, vol. 949, p. 012009.

FAN, GUİXİA, LİGUANG WANG, YİJUN CAO, AND CHAO Lİ., 2020. *Collecting Agent–Mineral Interactions in the Reverse Flotation of Iron Ore: A Brief Review*. Minerals. 10, no. 8: 681.

FERNÁNDEZ, Á., BELLA, J., & DORRONSORO, J. R., 2022. *Supervised outlier detection for classification and regression*. Neurocomputing. 486, 77–92.

GE, Z., 2017. *Review on data-driven modeling and monitoring for plant-wide industrial processes*. Chemometrics and Intelligent Laboratory Systems, 171(September), 16–25.

GOMEZ-FLORES, A., HEYES, G. W., ILYAS, S., & KİM, H., 2022. *Prediction of grade and recovery in flotation from physicochemical and operational aspects using machine learning models*. Minerals Engineering. 183(March), 107627.

JİAN, H., LİHUİ, C., & XİE, Y., 2020. *Design of Soft Sensor for Industrial Antimony Flotation Based on Deep CNN*. Proceedings of the 32nd Chinese Control and Decision Conference, CCDC 2020, 2492–2496.

Lİ, Z. MEİ, GUİ, W. HUA, & ZHU, J. YONG., 2019. *Fault detection in flotation processes based on deep learning and support vector machine.* Journal of Central South University, 26(9), 2504–2515.

MCCOY, J. T., & AURET, L., 2019. *Machine learning applications in minerals processing: A review. I*n Minerals Engineering (Vol. 132, pp. 95–109). Elsevier Ltd.

MONTANARES, M., GUAJARDO, S., AGUİLERA, I., & RİSSO, N., 2021. *Assessing machine learning-based approaches for silica concentration estimation in iron froth flotation.* 2021 IEEE International Conference on Automation/24th Congress of the Chilean Association of Automatic Control, ICA-ACCA 2021.

OLİVEİRA, E. M., 2017. *Quality Prediction in a Mining Process*. Retrieved September 7, 2019, from Kaggle.com website: https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process

POPLİ, K., AFACAN, A., LİU, Q., & PRASAD, V., 2018. *Development of online soft sensors and dynamic fundamental model-based process monitoring for complex sulfide ore flotation.* Minerals Engineering, 124(May), 10–27.

REN, L., WANG, T., LAİLİ, Y., & ZHANG, L., 2022. *A Data-Driven Self-Supervised LSTM-DeepFM Model for Industrial Soft Sensor.* IEEE Transactions on Industrial Informatics, 18(9), 5859–5869.

WEN, Z., ZHOU, C., PAN, J., NİE, T., ZHOU, C., & LU, Z., 2021. *Deep learning-based ash content prediction of coal flotation concentrate using convolutional neural network*. Minerals Engineering, 174, 107251

ZHANG, D., & GAO, X., 2021. S*oft sensor of flotation froth grade classification based on hybrid deep neural network.* International Journal of Production Research, 59(16), 4794–4810.

ZHANG, D., & GAO, X., 2022. *A digital twin dosing system for iron reverse flotation.* Journal of Manufacturing Systems, 63(March), 238–249. https://doi.org/10.1016/j.jmsy.2022.03.006

ZHANG, D., GAO, X., & Qİ, W., 2022. *Soft sensor of iron tailings grade based on froth image features for reverse flotation.* Transactions of the Institute of Measurement and Control, 44(15), 2928–2940.

ZHAO, B., HU, S., ZHAO, X., ZHOU, B., Lİ, J., HUANG, W., CHEN, G., WU, C., & LİU, K., 2022. *The application of machine learning models based on particles characteristics during coal slime flotation*. Advanced Powder Technology, 33(1), 103363.