Shahrood University of Technology

Journal of Mining and Environment (JME)

Journal homepage: www.jme.shahroodut.ac.ir

Iranian Society of Mining Engineering (IRSME)

Predicting open pit mine production using machine learning techniques: A case study in Peru

Marco Cotrina-Teatino[1]*, Jairo Marquina-Araujo[1], Eduardo Noriega-Vidal[1], José Mamani-Quispe[2], Johnny Ccatamayo-Barrios[3], Joe Gonzalez-Vasquez[4], and Solio Arango-Retamozo[4]

1. Department of Mining Engineering, Faculty of Engineering, National University of Trujillo, Trujillo, Peru
2. Department of Mining Engineering, University of Chile, Santiago, Chile
3. Department of Mining Engineering, National University of San Cristóbal de Huamanga, Ayacucho, Peru
4. Department of Industrial Engineering, National University of Trujillo, Trujillo, Peru

Article Info

Received 13 April 2024

Received in Revised form 1 May 2024

Accepted 15 May 2024

Published online 15 May 2024

DOI: 10.22044/jme.2024.14416.2703

Keywords

Machine learning
Open Pit Mine Production
Bayesian Regression
Predictive Modeling in Mining

Abstract

The primary objective of this research was to apply machine learning techniques to predict the production of an open pit mine in Peru. Four advanced techniques were employed: Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Bayesian Regression (RB). The methodology included the collection of 90 datasets over a three-month period, encompassing variables such as operational delays, operating hours, equipment utilization, the number of dump trucks used, and daily production. The data were allocated 70% for training and 30% for testing. The models were evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Variance Accounted For (VAF), and the Coefficient of Determination (R2). The results indicated that the Bayesian Regression model was the most effective in predicting production in the open pit mine. The RMSE, MAPE, VAF, and R2 for the models were 3686.60, 3581.82, 4576.61, and 3352.87; 12.65, 11.09, 15.31, and 11.90; 36.82, 40.72, 1.85, and 47.32; 0.37, 0.41, 0.41, and 0.47 for RF, XGBoost, KNN, and RB, respectively. This research highlights the efficacy of machine learning techniques in predicting mine production and recommends adjusting each model's parameters to further enhance outcomes, significantly contributing to strategic and operational management in the mining industry.

1. Introduction

Mining operations involve the continuous extraction and transport of significant volumes of material, using high-capacity equipment. Loading and haulage accounts for a substantial portion, approximately 50%, of the total operating costs in open pit mines [1]. Therefore, it is crucial that the truck loading and haulage system operates efficiently to optimize production and minimize energy consumption, thereby achieving mine performance targets [2] [3]. Production efficiency and throughput can vary significantly between mines [4] [5], depending on various factors such as the nature and planning of production processes, ore quality in the ore body, equipment availability and reliability, as well as challenges related to mine process design [6]. It is important to note that the performance of mining equipment is closely linked to its utilization, availability, and rated capacity [7] [8] [9].

Humans have the ability to learn from their daily experience thanks to their faculty of thought; for example, they can acquire knowledge through education or by reflecting on their thoughts and memories. In contrast, computers learn through algorithms, which is the foundation of machine learning (ML). ML employs computer algorithms to mimic the human learning process, allowing computers to identify and acquire real-world knowledge [10]. This, in turn, improves performance on specific tasks based on the newly

✉ *Corresponding author: mcotrinat@unitru.edu.pe (M. Cotrina-Teatino)*

acquired knowledge. According to the definition, "a computer program is said to learn from experience E corresponding to a certain class of tasks T and performance measure P, if the tasks T, are measured by P, improve with experience E" [11]. Although the initial concepts of ML emerged in the 1950s, it was not until the 1990s that it was consolidated as an independent field [10]. Machine learning algorithms are applied in a variety of fields, such as computer science [12] [13], health [14], environment [15], medicine [16], energy [17], engineering [18] [19] and services [20].

In open pit mining, the ability to forecast production is essential to plan operations efficiently and optimize the use of available resources [21]. In this context, the application of machine learning models has proven to be an invaluable resource to improve the accuracy in these projections [22]. Several models have been developed for this purpose, among which Random Forest, XGBoost, KNN and Bayesian Regression stand out [23]. In research related to production estimation, Baek and Choi [24] conducted a study to predict ore production and crusher utilization in a subway limestone mine using a neural network with five hidden layers and 300 neurons in each hidden layer. In Baek and Choi [25], two ANN models were constructed for the morning and afternoon haulage shifts, respectively. According to the study, the MAPE for morning and afternoon was 4.78% and 5.26%, respectively, with a coefficient of determination of 0.99 each. In addition, Choi *et al.* [26], used machine learning models including ANN, Support Vector Machine (SVM), Random Forests (RF), Classification and Regression Tree (CART) and K-Nearest Neighbors to estimate ore production in an open pit limestone mine in South Korea. Among the models, the SVM algorithm obtained better results with the highest accuracy. In addition, Nartey *et al.* [27] applied four machine learning algorithms, including ANN, Random Forest, GBR and DT, to forecast the ore production in an open pit mine. The results showed that ANN achieved the highest accuracy, with an R^2 of 0.8003 and a MAPE of 4.23%.

According to the literature consulted, there is a scarcity of applications of machine learning models to predict production in open pit mines, with no records of such applications in the Peruvian context. This lack makes it difficult to identify the most effective strategy to forecast mine production using machine learning algorithms. Therefore, the main objective of this study is to apply four different machine learning models to predict production in open pit mines. The methods evaluated include Random Forest (RF), XGBoost, KNN and Bayesian Regression. The various models were evaluated and compared using metrics such as coefficient of determination (R^2), mean absolute percentage error (MAPE), variance accounted for (VAF), root mean square error (RMSE) and correlation coefficient (R). This article is intended to serve as a guide for future research on the use of machine learning in mining production modeling and prediction. The main contributions of this study include the application of machine learning techniques for the prediction of open pit mining production, a novel approach in this field, providing valuable information for the optimization of mining operations. The structure of the article is as follows: Section 2 describes the methodology used in the research. Section 3 presents the results obtained and the corresponding discussions are carried out. Section 4 details the conclusions reached in the study and, finally, a list of bibliographical references used in the research is provided.

2. Methodology

2.1. Description of the mine

The mine under study is located near Huamachuco, in the northern highlands of the La Libertad region of Peru. This important gold deposit produces mainly gold, with copper as a by-product, contributing substantially to regional mining production. Located at an altitude of 3800 meters above sea level, the facility employs advanced drilling and blasting techniques to effectively fragment the rock mass for further processing. Once fragmented, the ore is systematically loaded into a fleet of 90 dump trucks, each with a capacity of 26 m^3. These trucks are meticulously organized to operate in continuous day and night shifts, ensuring uninterrupted transport of both ore and waste rock, integral to the mining cycle that includes mining, drilling, blasting, loading, and hauling. The equipment portfolio is comprehensive, with state-of-the-art hydraulic excavators, drilling rigs and blast hole drills, all deployed to improve the efficiency and productivity of mining operations. Total resources (measured + indicated + inferred) are at approximately 1080000 ounces of gold by 2024.

2.2. Data analysis methods

The analysis methods employed in this research integrate statistical techniques with advanced machine learning algorithms. Initially, a descriptive analysis of the data was performed, using summarized statistics to understand the basic characteristics of the operational variables. Subsequently, a correlation analysis was applied using Pearson correlation coefficients to identify linear relationships between the variables. This step was crucial for exploring how variables such as the number of dump trucks and excavator operating hours are related to daily production. For predictive modeling, four machine learning techniques were selected: Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Bayesian Regression (RB). Each model was trained and evaluated using a dataset split into 70% for training and 30% for testing, assessing their performance through metrics such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Variability Acceptance Factor (VAF), and the Coefficient of Determination (R^2).

All analyses were performed using Python version 11.7, with Jupyter Notebook as the web-based interactive development environment. This setup facilitated efficient handling and processing of data, enabling rigorous modeling and precise performance evaluation.

2.3. Description of the database

The database used in this analysis comprises 90 historical data sets collected from mine X over a period of 90 days (equivalent to 3 months December 2023 to February 2024). These datasets include a number of parameters, namely: average daily truck (dump trucks), average excavator operating hours, average excavator operating delay hours, average equipment usage percentage, and daily production (m^3 per day). Since the mine operates in two daily shifts, each row of data reflects the average values for both shifts.

2.4. Machine learning model

2.4.1. Random Forest (RF)

Random Forest (RF) is an ensemble technique in which the results of a collection of random decision trees are combined to obtain an overall prediction. This method is applicable for both regression and classification of practical interest. At each internal node of the tree, a binary decision is made based on a Boolean test [28]. For example, if the attribute selected for splitting is ordinal, the test involves determining whether the attribute value is above a threshold. Instances for which the answer to the test is true (i.e., the value of the attribute is above the specified threshold) are assigned to one of the child nodes. Those with a false response (i.e., the attribute value is below the threshold) are assigned to the other child node. In this way, the training data are divided into separate subsets. The splitting is done so that, within each subset, the classes are better separated (in classification problems) or the prediction error is minimized (in regression). In random trees, the Boolean test of a specific internal node is selected as the best split resulting from considering only a randomly chosen subset of attributes [29]. The tree grows until a new split does not lead to purer nodes or a specified pre-pruning criterion is met (e.g., there are too few instances assigned to a node, or the maximum depth of the tree is reached). Each tree in the forest is constructed from a Bootstrap sample independent of the data, as in bagging [30].

Predictions will be made at the terminal nodes (leaves) of the tree based on the training instances that have been assigned to those nodes by the testing sequence at the root node and subsequent intermediate nodes connecting the root to the corresponding leaf. In regression, the prediction is the average value of the response variable over the training instances assigned to that leaf. In classification, the final ensemble prediction is obtained by majority vote. In regression, the outcome of an ensemble of size T is the average of the predictions of the random trees in the ensemble.

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)}(x) \tag{1}$$

The advantages of Random Forest are that it handles large data sets well, offering robustness against overfitting and providing insights into the importance of variables, and the disadvantages are that it can be computationally demanding, leading to longer training times, and may have difficulties with extrapolations outside the range of the training data [30].

2.4.2. Extreme Gradient Boosting (XGBoost)

XGBoost is a parallel tree boosting system based on the gradient boosting method [31] [32]. It uses a model composed of a set of classification and regression trees (CART) [33]. Although XGBoost appears similar to GBDT, it has some inherent features that differ from GBDT, such as the second-order Taylor expansion and the built-in normalization function. The XGBoost model can be briefly explained as follows:

For a data set $D = (x_i, y_i)(x_i \in R^m, y_i \in R, i = 1,2, \ldots, n)$ containing n instances with m

dimensions, and a model trained with G trees, the predictions are obtained by the following formula:

$$\hat{y}_i = \sum_{k=1}^{G} f(x_i), f_k \in F(i = 1,2, \dots, n) \qquad (2)$$

Where f is the hypothesis space, and $f(x)$ is a regression tree: $F = \{f(x) = w_{q(x)}\}(q{:}\, R^m \rightarrow \{1,2, \dots, T\}, w \in R^T)\ q(x)$. Here, $q(x)$ is the leaf node, and w is the leaf score [34]. To construct an ideal model, it is necessary to minimize the objective function to find the optimal parameters. This can be divided into a loss function (L) and a model complexity function (Ω).

$$L = \sum_{i=1}^{n} L(y_i - \hat{y}_i)^2 \qquad (3)$$

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (4)$$

The advantage of XGBoost is that it is efficient and performs well on various data types, with regularization methods that help prevent overfitting. The disadvantage is its complexity can make hyperparameter tuning challenging and less intuitive to understand the relationships between features [33].

2.4.3. K-Nearest Neighbors (KNN)

The nearest neighbor algorithm, also known as KNN or k-NN, is a nonparametric supervised learning classifier that uses proximity to perform classifications or predictions based on the clustering of a single data point [35]. The distance metric in nearest neighbor methods is the simple Euclidean distance. That is, the distance between two patterns $(x_{11}, x_{12}, \dots, x_{1n})$ and $(x_{21}, x_{22}, \dots, x_{2n})$ is calculated using the following formula:

$$DE = \sqrt{\sum\nolimits_{j=1}^{n} (x_{1j} - x_{2j})^2} \qquad (5)$$

This approach considers the entire training set as the search space. When a test instance is presented, the distance between this instance and all points in the training set is calculated. Then, the k points closest to the test instance are selected, where k is a user-specified parameter. The class of the test instance is determined by majority voting the classes of the k nearest neighbors. In the regression case, the prediction is performed by taking the average of the target values of the k nearest neighbors [35].

The advantage of KNN is that it is simple and adaptive, effective in recommender systems where similarity between items is crucial. The disadvantage is that it is sensitive to data scale and irrelevant features, and computationally inefficient for large datasets [35].

2.4.4. Bayesian regression (RB)

Bayesian regression, in particular, is a powerful approach that uses Bayes' theorem to update beliefs about model parameters as new data are observed. In Bayesian regression, the goal is to estimate the posterior distribution of model parameters given a set of observed data. This is achieved by combining the likelihood of the data and the prior distribution of the parameters using Bayes' theorem [36]. The posterior distribution, which represents the updated knowledge about the model parameters after observing the data, is calculated as:

$$P(\theta|D) = \frac{P(D|\theta)\, P(\theta)}{P(D)} \qquad (6)$$

Where $P(\theta \mid D)$ is the posterior distribution of the parameters given the data set $P(D \mid \theta)$ is the likelihood of the data given the parameter vector $P(\theta)$ is the prior distribution of the parameters and $P(D)$ is the marginal likelihood of the data. In Bayesian regression, the model can include a wide range of prior distributions for the parameters, allowing prior information or expert knowledge to be incorporated into the inference process [36].

The advantage of RB is that it incorporates prior knowledge, improving predictions in scenarios with uncertain data and provides uncertainty estimates. The disadvantage is that it depends on the correct choice of a priori distribution, may bias the results, and is computationally intensive [36].

2.5. Evaluation metrics

The RMSE should be zero for a perfect model. The RMSE of a model prediction relative to the observed values is defined as the square root of the mean square error [37]:

$$RMSE = \sqrt{\frac{1}{n_{ts}} \sum_{i} (o_i - p_i)^2} \qquad (7)$$

The Mean Absolute Percentage Error (MAPE) is a commonly used metric to assess the accuracy of a prediction model relative to the observed values. It is calculated as the average of the absolute value of the individual percentage errors between the predictions (p_i) and the observed values (o_i), expressed as a percentage of the true value [38]:

$$MAPE = \frac{1}{n_{ts}}\sum_{i}\left|\frac{o_i - p_i}{o_i}\right| x\,100 \qquad (8)$$

The Variability Acceptance Factor (VAF) provides a measure of how much variability in the data is explained by the model relative to the total variability present in the observed data. It is calculated as the ratio of the sum of the squares of the differences between the observed values (o_i) and the predicted values (p_i) to the total sum of the squares of the differences between the observed values and their mean $(o_i - \bar{o})^2$ [38]:

$$VAF = (1 - \frac{\sum_i (o_i - p_i)^2}{\sum_i (o_i - \bar{o})^2} x\,100 \qquad (9)$$

The R-Squared value (known as the coefficient of determination) describes how much of the variance between the two variables (observed and predicted values) describes the legal fit. This can be determined as [38]:

$$R^2 = \frac{(\sum(o_i - \bar{o_i})(p_i - \bar{p_i}))^2}{\sum(o_i - \bar{o_i})^2 \sum(p_i - \bar{p_i})^2} \qquad (8)$$

3. Results and discussions

Table 1 presents a statistical summary of the database, including the count, average, standard deviation (std), minimum, percentiles (25%, 50% and 75%) and maximum values for each of the parameters such as operational delays, operational hours, usage, number of dump trucks and production in m^3/day.

Table 1. Statistical description of the database

(p-value)	Operational delays (h)	Operational hours (h)	Usage (%)	N° of dump trucks	Production (m^3/day)
Count	90	90	90	90	90
Mean	4.45	14.09	49.02	24.19	26253.19
Std	1.77	2.29	8.14	4.94	5759.70
Min	2.21	8.78	24.00	14.00	10377.29
25%	3.38	12.85	45.00	21.00	23229.15
50%	4.12	13.98	49.00	23.00	26430.18
75%	4.94	15.66	54.00	27.75	29849.17
Max	15.22	20.15	68.00	36.00	40795.63

Table 2 shows the correlation matrix of the database, where each cell contains the Pearson correlation coefficient. There is a moderate positive correlation between the number of dump trucks and daily production (0.45). Likewise, a significant negative correlation is identified between equipment usage and daily production (-0.25). In addition, a positive correlation is observed between excavator operating hours and daily production (0.50), and a weak negative correlation between excavator operating delays and daily production (-0.25).

Table 2. Correlation matrix of the database

	Operational delays (h)	Operational hours (h)	Usage (%)	N° of dump trucks	Production (m^3/day)
Operational delays (h)	1.00				
Operational hours (h)	-0.22	1.00			
Usage (%)	-0.46	0.39	1.00		
N° of dump trucks	0.18	0.38	0.16	1.00	
Production (m^3/day)	-0.25	0.50	0.40	0.45	1.00

In the research work a rigorous partitioning of the data has been applied, reserving 30% of the data for testing and 70% for training, as shown in Table 3. It summarizes the statistics of the data distributed for training and testing of each artificial intelligence model, where the inputs: operating delays, usage (%), operating hours and number of dump trucks and the output is the daily production. It can be observed that for training the data is a total of 63 for each variable and for test it is 27 data of the total.

Table 3. Statistics of distributed data for training and testing

	Input								Output	
	Operational delays (h)		Usage (%)		Operational hours (h)		N° of dump trucks		Producción (m^3/día)	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Quantity	63.00	27.00	63.00	27.00	63.00	27.00	63.00	27.00	63.00	27.00
Mean	4.29	4.82	7.70	50.41	14.06	14.14	39.30	23.93	60273.16	26363.95
Std	1.35	2.49	6.87	10.21	2.32	2.22	4.77	5.38	14237.67	4707.60
10%	2.21	2.53	0.00	24.00	8.95	8.78	29.00	15.00	23867.76	12891.83
25%	3.42	3.31	0.50	47.00	12.55	13.14	37.00	20.50	53298.33	24558.60
50%	4.10	4.30	9.05	51.00	14.03	13.80	38.00	22.00	59420.66	27959.76
75%	4.74	5.07	12.71	56.50	15.22	15.69	43.00	26.50	69613.64	29217.99
Max	10.31	15.22	23.95	68.00	20.15	17.64	51.00	35.00	93829.95	33247.40

Table 4 presents the hyperparameters used in the machine learning models. For Random Forests, 300 estimators were used with a maximum depth of 30 for each tree. In addition, a minimum split of 2 samples was used at each node and at least one sample was required to form a leaf. In Extreme Gradient Boosting, the learning rate (Eta) was set to 0.05 to control the learning rate, while the maximum tree depth was limited to 6. For KNN, 5 nearest neighbors with uniform weights were considered and a Minkowski distance metric was used with an additional parameter of 2. For Bayesian regression, 500 iterations were performed with a convergence tolerance of 1x10^{-4}. The precision and scaling parameters for the a priori distributions were also set to 1x10^{-7}.

Table 4. Hyperparameters of the machine learning models used

Method	Parameter	Value	Description
Random Forests (RFs)	N° de estimadores	300	Number of trees
	Max_depth	30	Maximum depth of each tree
	Min_samples_split	2	Min. samples required to split a node
	Min_samples_leaf	1	Min. samples required to be a leaf
	Max_features	'auto'	Number of features to find the best split
	Bootstrap	True	Use of samples with replacement
	Random_state	42	Seed for randomization
	Verbose	0	Verbosity in training
	Obb_score	False	Use out-of-bag samples to estimate accuracy
Extreme Gradient Boosting (XGBoost)	Booster	Gbtree	Defines the type of base model used in the algorithm
	Eta (learning rate)	0.05	Controls learning speed
	Max_depth	6	Limits the complexity of each tree
	Min_child_weight	1	Controls tree growth
	Gamma	0.0	Regulates the creation of new leaves in the tree
	Subsample	1.0	Controls the fraction of training instances used in each tree
	Colsample_bytree	0.6	Controls the fraction of features used in each tree
	Eval_metric	‘rmse’	Defines the model evaluation metric
	Estimators	1000	Specifies the number of trees in the ensemble
	Objective	Reg: squared error	Defines the optimized loss function
	Random state	42	Controls the reproducibility of the results
	Scale_pos_weight	1	Used in unbalanced classification problems.
K-Nearest Neighbors	N_neighbors	5	Specifies the number of nearest neighbors considered.
	Weights	‘uniform’	Controls how neighbor contributions are weighted.
	Algorithm	‘auto’	Defines the method used to calculate the nearest neighbors.
	Metric	‘minkowski’	Specifies the distance measure used.
	‘p’	2	Additional parameter to calculate the distance in the Minkowski algorithm.
Bayessian Ridge (RB)	N_iter	500	Number of iterations to perform parameter estimation.
	tol	1e-4	Tolerance for algorithm convergence.
	Alpha_1	1e-7	Precision parameter for a priori distribution of weights.
	Alpha_2	1e-7	Precision parameter for a priori distribution of error variance.
	Lambda_1	1e-7	Scale parameter for the a priori distribution of the weights.
	Lambda_2	1e-7	Scale parameter for the a priori distribution of the error variance.
	Compute_score	True	Indicates whether or not to calculate the score during model fitting.
	Fit_intercept	True	Indicates whether or not to fit the model intercept.

Hyperparameters are very important for prediction results. Baek and Choi [24] in their hyperparameters of their neural network used was 5 hidden layers and each layer had 300 neurons. Likewise, Nartey [27] in his machine learning models (Random Forest) used a minimum sample Split of 1-5, number of estimators of 1-500, step size of 10, maximum Depth of 1-8, this performed in order to obtain the optimal hyperparameters, in the research the parameters for random forest was 300 estimators, with a depth of 30, minimum samples of 2, max_features: auto and a randomness seed of 42.

3.1. Test and training results of the machine learning models

Figure 1 presents a comparative analysis between predicted and actual values obtained using the random forest (RF) model, separated by training and test data sets (unseen data). In the training data set, a correlation of 0.88 was obtained, which is consistent with the tendency of the models to learn specifically from the data they are trained on. Likewise, in the test data set a correlation of 0.63 was achieved, indicating a considerable fit between the model predictions and the actual production quantity.

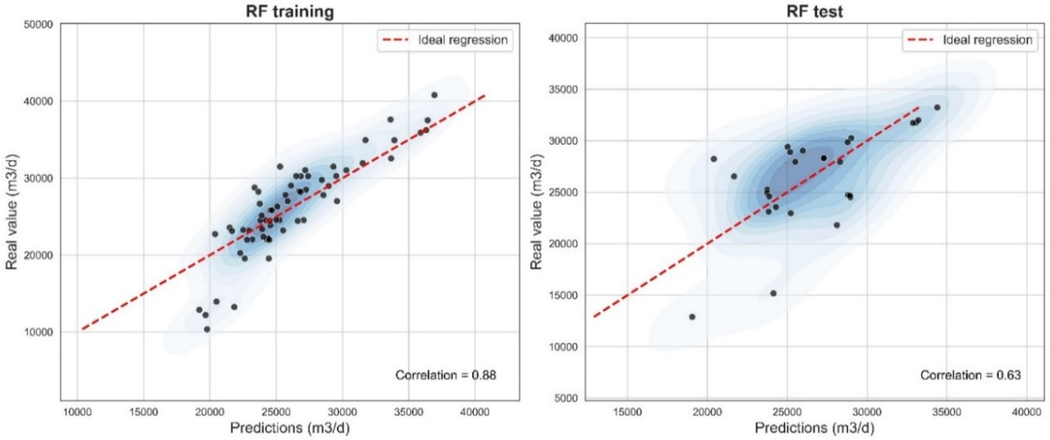

Figure 1. Comparison of actual vs. predicted values using the Random Forest (RF) model

Figure 2 shows a comparative analysis of the predicted results versus the actual values of the XGBoost model. A perfect correlation (1.00) is observed in the training set, indicating a high level of accuracy in that set. However, the accuracy of the test data set was 0.68.

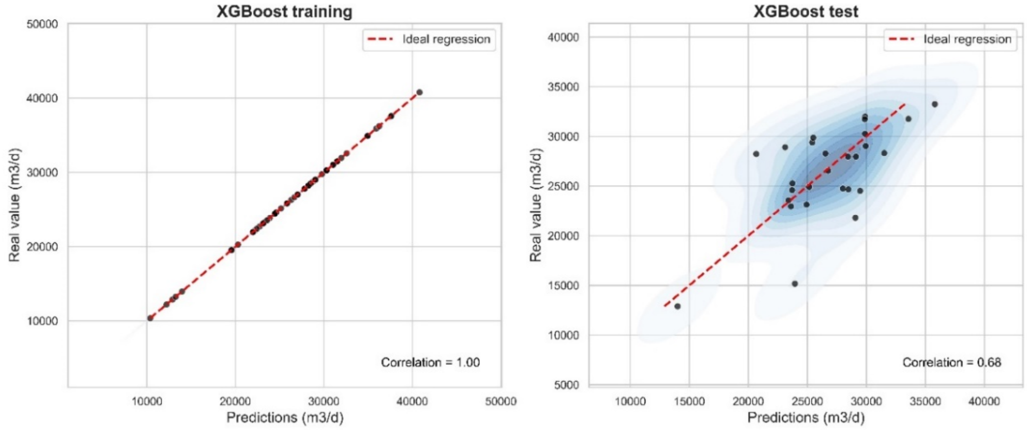

Figure 2. Comparison of actual vs. predicted values using the XGBoost model.

Figure 3 shows the relationship between the values predicted by the KNN (K-Nearest Neighbors) model and the actual values. With a correlation of 0.67 in the training data set, indicating a high level of accuracy. A correlation of 0.67 is obtained in the test data sets, showing a good ability for predictions on test data (unseen data).

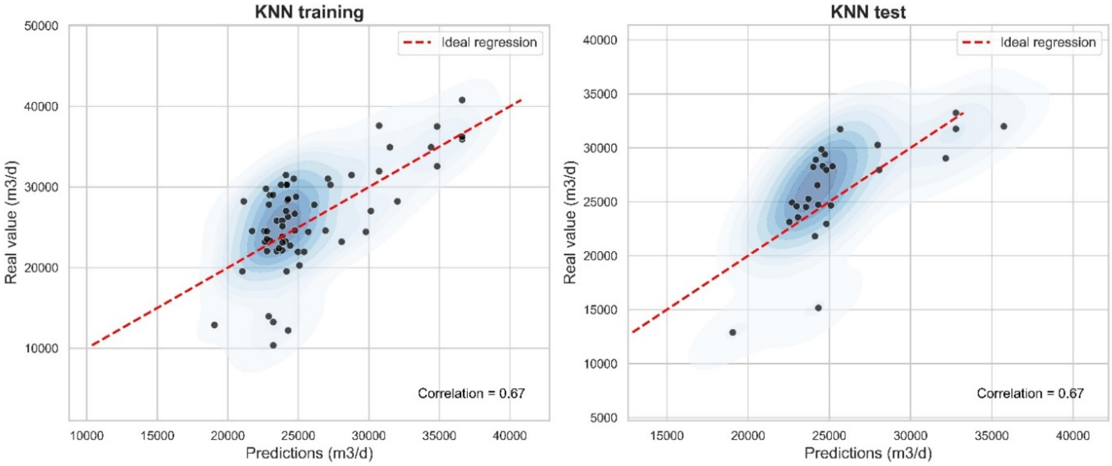

Figure 3. Comparison of actual vs. predicted values using the KNN model.

Figure 4 illustrates the predictive ability of the Bayesian Regression (RB) model by comparing the predicted values with the actual values on the training and test data sets. The model achieves a correlation of 0.60 in the training set, 0.69 in the test set. This demonstrates consistency in predicting daily production in cubic meters at the mining company.

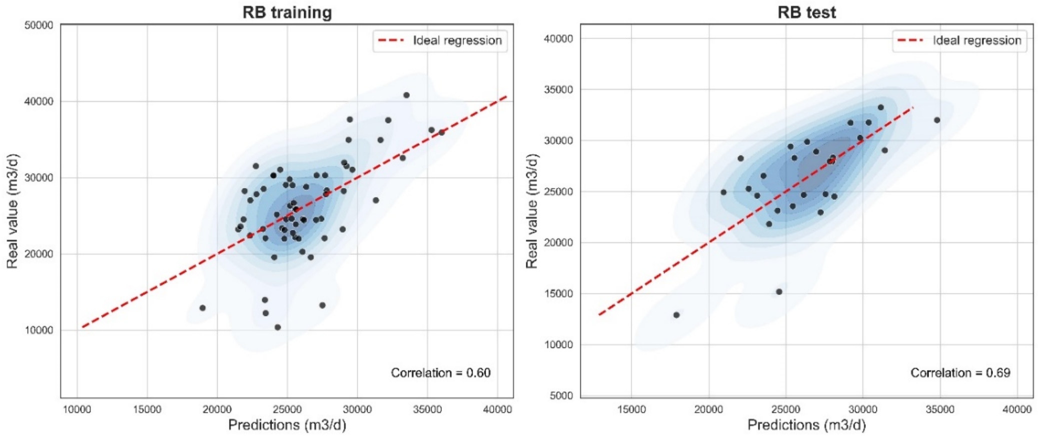

Figure 4. Comparison of actual vs. predicted values using the Bayessian Ridge (RB) model.

3.2. Comparison and evaluation of the machine learning models.

For comparison purposes, Figure 5 and Figure 6 show the prediction results of the machine learning models on an independent test data set. In addition, Table 5 presents the prediction results, in which the proposed RB model yielded an RMSE value of 3352.87 and an R^2 of 0.47, while the Random Forest (RF) model yielded an RMSE of 3686.00 and an R^2 of 0.37. The XGBoost model yielded an RMSE of 3581.82 with an R^2 of 0.41, finally the KNN model yielded an RMSE of 4576.61 and the correlation coefficient was an R^2 of 0.41.

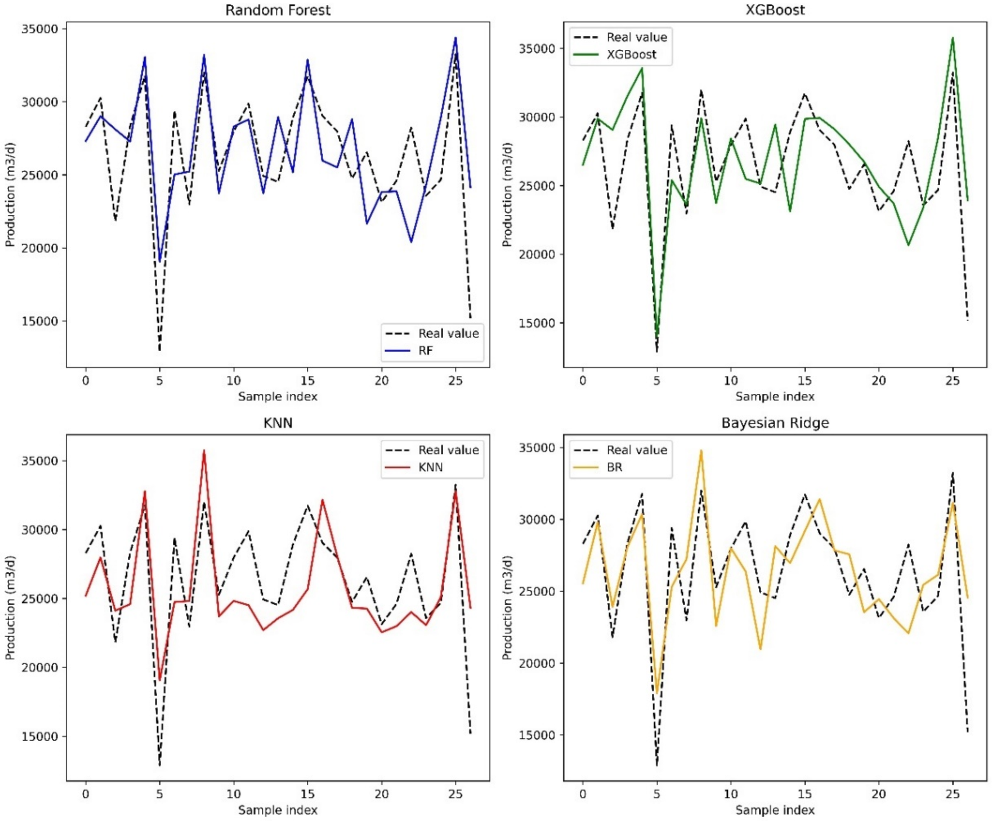

Figure 5. Model prediction results on a test data set.

Table 5. Performance of Machine Learning Methods.

Metric	RF	XGBoost	KNN	RB
RMSE	3686.60	3581.82	4576.61	3352.87
MAPE	12.65	11.09	15.31	11.90
VAF	36.82	40.72	1.85	47.32
R^2	0.37	0.41	0.41	0.47
R	0.63	0.68	0.67	0.70

Baek and Choi [25] obtained a MAPE of 4.78% in the morning and 5.26% in the afternoon, with a coefficient of determination (R^2) of 0.99 each. Likewise, Nartey [27] applied 4 machine learning algorithms where the results shown by ANN reached an R^2 of 0.8003 and a MAPE of 4.23%. Comparing these results with those obtained in which the best model was Bayesian Regression with a MAPE of 11.90%, an R^2 of 0.47 and a correlation of 0.70. This indicates that the hyperparameters of each model need to be improved; this is also due to the distribution of the information obtained from the field.

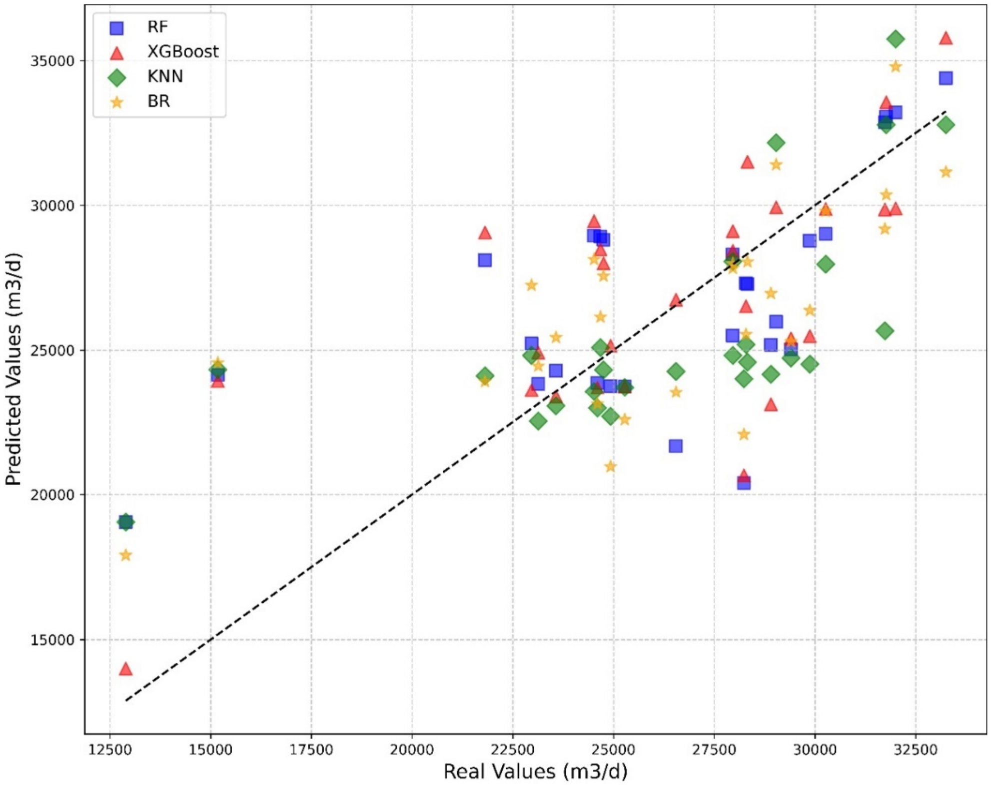

Figure 6. Actual production (m^3/day) versus predicted production (m^3/day) in the test data set for each machine learning model.

The separate test data set, reserved for evaluating the performance of the machine learning models, comprised 27 data points and reflected the operational diversity of the mining process. As shown in Table 3, the test data exhibited an average of 4.82 hours in operational delays, an average utilization rate of 50.41%, and 14.14 operational hours, with an average of 23.93 dump trucks in operation. Daily production output in the test data averaged 26,363.95 m³/day. The standard deviation for production was 4,707.60 m³/day, indicating variability in daily volumes. Production data ranged from a minimum of 12,891.83 m³/day to a maximum of 33,247.40 m³/day, with quartiles positioned at 24,558.60, 27,959.76 and 29,217.99 m³/day, respectively. These statistics underscore the heterogeneity of the test data and validate the robust predictive capabilities of the Bayesian Regression model, which achieved close alignment with actual production values.

4. Conclusions

This research evaluated the performance of machine learning algorithms: RF, XGBoost, KNN, and RB, for predicting daily production at an open pit mine in Peru. For this purpose, 90 data sets were utilized. Input parameters included the average daily number of dump trucks, the percentage of excavator utilization, the average daily number of operating hours and delays, with daily production serving as the output parameter. Among the implemented models, Bayesian regression proved to be the most efficient in prediction, achieving a Coefficient of Determination (R²) of 0.47, a Mean Absolute Percentage Error (MAPE) of 11.90%, and a Variance Accounted For (VAF) of 47.32%, followed by KNN, XGBoost, and finally the Random Forest model. The results enable the prediction of mine production with moderate accuracy using the Bayesian regression model. Overall, the study has demonstrated that machine learning techniques can be relevant for modeling and predicting the production of an open pit mine.

It is essential to highlight that this study contributes to originality and literature by exploring the less common use of Bayesian regression in this context and addresses the risk level of the study area by recommending the inclusion of more variables such as mechanical availability, operational efficiencies, and cycle times for future work. The use of additional machine learning models is also suggested to enhance the prediction of mine production. These recommendations are crucial for guiding future research and adapting operational strategies in the mining sector.

Referencias

[1]. Alarie, S. and Gamache, M. (2002). Overview of solution strategies used in truck dispatching systems for open pit mines. *Int J Surf Min Reclamat Environ, 16*(1), 59-76.

[2]. Arteaga, F., Nehring, M. and Knights, P. (2018). The equipment utilization versus mining rate trade-off in open pit mining. *Int J Min Reclamat Environ*, *32*(7), 495-518.

[3]. Santelices, G., Pascual, R., Luer, A., Mac Cawley, A. and Galar, D. (2017). Integrating mining loading and hauling equipment selection and replacement decisions using stochastic linear programming. *Int J Min Reclamat Environ*, *31*(1), 52-65.

[4]. Edwards, D., Holt, G. and Harris, F. (2002). Predicting downtime costs of tracked hydraulic excavators operating in the UK opencast mining industry. *Construct Manag Econ*, *20*(7), 581-591.

[5]. Fisonga, M. and Mutambo, V. (2017). Optimization of the fleet per shovel productivity in surface mining: case study of Chilanga Cement, Lusaka Zambia. *Cogent Eng*, *4*(1), 1386852.

[6]. Ozdemir, B. and Kumral, M. (2018). Appraising production targets through agent-based Petri net simulation of material handling systems in open pit mines. *Simulat Model Pract Theor*, *87*, 138-154.

[7]. Lanke, A. Hoseinie, S. and Ghodrati, B. (2016). Mine production index (MPI)-extension of OEE for bottleneck detection in mining. *Int J Min Sci Technol*, *26*(5), 753-760.

[8]. Soofastaei, A. Karimpour, E. Knights, P. and Kizil, M. (2018). Energy-efficient loading and hauling operations. *Green Energy Technol*, 121-146.

[9]. Kaba, F. Temeng, V. and Eshun, P. (2016). Application of Discrete event simulation in mine production forecast. *Ghana Min J*, *16*(1).

[10]. Jung, D. and Choi, Y. (2021). Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. *Minerals*, *11*(2), 148.

[11]. Michalski, R. Carbonell, J. and Mitchell, T. (2013). Machine Learning: An Artificial Intelligence Approach. *Springer Science & Business Media: Berlin/Heidelberg.*

[12]. Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing*, *27*, 504-518.

[13]. Handelman, G. Kuan, H. Chandra, R. Razavi, A. Huang, S. Brooks, M. Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *Am. J. Roentgenol*, *212*(1), 38-43.

[14]. Spasic, I. and Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, *8*(3).

[15]. Bellinger, C. Mohomed, M. Zaiane, O. and Osornio, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(907).

[16]. Senders, J. Staples, P. Karhade, A. Zaki, M. Gormley, W. Broekman, M. and Arnaout, O. (2018). Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurgery*, *109*, 476-786.

[17]. Mosavi, A. Salimi, M. Faizollahzadeh, S. Rabczuk, T. Shamshirband, S. and Varkonyi, A. (2019). State of the Art of Machine Learning Models in Energy Systems, a Systematic Review. *Energies*, *12*(7), 1301.

[18]. Jenis, J. Ondriga, J. Hrcek, S. Brumercik, F. Cuchor, M. and Sadovsky, E. (2023). Engineering Applications of Artificial Intelligence in Mechanical Design and Optimization. *Machines*, *11*(6), 577.

[19]. Whitehall, B. and Lu, S. (1991). Machine learning in engineering automation —The present and the future. *Computers in Industry*, *17*(2-3), 91-100.

[20]. Portugal, I. Alencar, P. and Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227.

[21]. Mendoza, J. (2021). Optimización del valor presente neto aplicando secuenciamiento de fases direccionadas en el diseño del pit del proyecto Cotabambas-Panoro Minerals. Arequipa.

[22]. Huang G, G. Y. (2023). Application of Machine Learning in Material Synthesis and Property Prediction. *Materials (Basel), 16*(17), 5977.

[23]. Sarker, I. H. (2021). Machine Learning: Algorithms, Real World Applications and Research Directions. *SN Computer Science*, *2*, 160.

[24]. Baek, J. and Choi, Y. (2019). Deep neural network for ore production and crusher utilization prediction of truck haulage system in underground mine. *Appl Sci, 9*(19), 4180.

[25]. Baek, J. and Choi, Y. (2020). Deep neural network for predicting ore production by truck-haulage systems in open-pit mines. *Appl Sci*, *10*(5), 1657.

[26]. Choi, Y. Nguyen, H. Bui, X. Nguyen-Thoi, T. and Park, S. (2021). Estimating ore production in open-pit mines using various machine learning algorithms based on a truck-haulage system and support of internet of things. *Nat Resour Res*, *30*, 1141-1173.

[27]. Nartey, F. Kwasi, A. Nkrumah, M. and Kweku, C. (2024). Predicting open-pit mine production using machine learning techniques. *Journal of Sustainable Mining*, 23(2).

[28]. Fenández-Delgado, M. Cernadas, E. Barro, S. and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res*, *15*, 3133-3181.

[29]. Ho, T. (1998). The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. *Intell, 20*(8), 832-844.

[30]. Breiman, L. (1996). Bagging predictors. *Mach. Learn*, *24*(2), 123-140.

[31]. Friedman, J. (2001). Greedy boosting approximation: A gradient boosting machine. *Annal. Stat*, *29*(5), 1189-1232.

[32]. Nabavi, Z. Mirzehi, M. Dehghani, H. and Ashtari, P. (2023). A Hybrid Model for Back-Break Prediction using XGBoost Machine learning and Metaheuristic Algorithms in Chadormalu Iron Mine. *Journal of Mining and Environment*, *14*(2), 689-712.

[33]. Breiman, L. Friedman, J. and Olshen, R. (1984). Classification and regression trees. *wadsworth int, Group*, *37*(15), 237-251.

[34]. Emrah, U. Dagasan, Y. and Topal, E. (2021). Mineral grade estimation using gradient boosting regression trees. *International Journal of Mining*, *Reclamation and Environment, 35*(10), 728-742.

[35]. Yu, K. and Zhang, X. (2002). Kernel Nearest Neighbor Algorithm. *Neural Processing Letters*, *15*, 147-156.

[36]. Bárcena, M. J. Garín , M. A. and Matrín, A. (2017). Un simulador para asistir en la enseñanza del teorema de Bayes.

[37]. Patel, A. Chatterjee, S. and Gorai, A. (2019). Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades. *Earth Sci Inform*, *12*, 197-210.

[38]. Prasad, K. Gorai, A. and Goyal, P. (2016). Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmos Environ*, *128*, 246-262.

پیش‌بینی تولید معدن روباز با استفاده از تکنیک‌های یادگیری ماشین: مطالعه موردی در پرو

مارکو کوترینا-تاتینو [1*]، ژایرو مارکینا-آراوخو[1]، ادواردو نوریگا-ویدال[1]، خوزه مامانی-کوئیسپ[2]، جانی ککاتامایو-باریوس[3]، جو گونزالس-واسکز[4]، و سولیو آرانگو-رتاموزو[4]

[1] .گروه مهندسی معدن، دانشکده مهندسی، دانشگاه ملی تروخیلو، تروخیلو، پرو

[2] .گروه مهندسی معدن، دانشگاه شیلی، سانتیاگو، شیلی

[3] .گروه مهندسی معدن، دانشگاه ملی سن کریستوبال د هوامانگا، ایاکوچو، پرو

[4]. گروه مهندسی صنایع، دانشگاه ملی تروخیلو، تروخیلو، پرو

ارسال ۲۰۲۴/۰۴/۱۳، پذیرش ۲۰۲۴/۰۵/۱۵

* نویسنده مسئول مکاتبات: mcotrinat@unitru.edu.pe

چکیده:

هدف اصلی این تحقیق استفاده از تکنیک‌های یادگیری ماشین برای پیش‌بینی تولید یک معدن روباز در پرو بود. چهار تکنیک پیشرفته مورد استفاده قرار گرفت: جنگل تصادفی (RF)، تقویت گرادیان شدید (XGBoost)، نزدیکترین همسایگان K (KNN)، و رگرسیون بیزی (RB). این روش شامل مجموعه‌ای از ۹۰ مجموعه داده در یک دوره سه ماهه بود که متغیرهایی مانند تاخیرهای عملیاتی، ساعات کار، استفاده از تجهیزات، تعداد کامیون‌های کمپرسی مورد استفاده و تولید روزانه را در بر می‌گرفت. داده ها ۷۰ درصد برای آموزش و ۳۰ درصد برای آزمون اختصاص داده شد. مدل‌ها با استفاده از معیارهایی مانند ریشه میانگین مربعات خطا (RMSE)، میانگین درصد مطلق خطا (MAPE)، واریانس محاسبه شده برای (VAF) و ضریب تعیین (R2) ارزیابی شدند. نتایج نشان داد که مدل رگرسیون بیزی بیشترین تاثیر را در پیش بینی تولید در معدن روباز دارد. RMSE، MAPE، VAF و R2 برای مدل‌ها ۳۶۸۶.۶۰، ۳۵۸۱.۸۲، ۴۵۷۶.۶۱ و ۳۳۵۲.۸۷ بود. ۱۲.۶۵، ۱۱.۰۹، ۱۵.۳۱ و ۱۱.۹۰; ۳۶.۸۲، ۴۰.۷۲، ۱.۸۵ و ۴۷.۳۲; ۰.۳۷، ۰.۴۱، ۰.۴۱، و ۰.۴۷ برای RF، XGBoost، KNN و RB به ترتیب. این تحقیق کارایی تکنیک‌های یادگیری ماشین را در پیش‌بینی تولید معدن برجسته می‌کند و توصیه می‌کند که پارامترهای هر مدل را برای بهبود بیشتر نتایج تنظیم کنید، که به طور قابل‌توجهی به مدیریت استراتژیک و عملیاتی در صنعت معدن کمک می‌کند.

کلمات کلیدی: یادگیری ماشینی، تولید معدن روباز، رگرسیون بیزی، مدل سازی پیش بینی در معدن.