 doi:10.20944/preprints202102.0326.v1

Article

A comparative study on supervised machine learning algorithms for copper recovery quality prediction in a leaching process

Victor Flores [1,*] and Claudio Leiva [2]

[1] Department of Computer and Systems Engineering, Universidad Católica del Norte, 1270709 Antofagasta, Chile; vflores@ucn.cl
[2] Department of Chemical Engineering Universidad Católica del Norte, 1270709 Antofagasta, Chile; cleiva01@ucn.cl
* Correspondence: Víctor Flores, vflores@ucn.cl

Abstract: The copper mining industry is increasingly using artificial intelligence methods to improve cop-per production processes. Recent studies reveal the use of algorithms such as Artificial Neural Network, Support Vector Machine, and Random Forest, among others, to develop models for predicting product quality. Other studies compare the predictive models developed with these machine learning algorithms in the mining industry, as a whole. However, not many copper mining studies published compare the results of machine learning techniques for copper recovery prediction. This study makes a detailed comparison between three models for predicting copper recovery by leaching, using four datasets resulting from mining operations in northern Chile. The algorithms used for developing the models were Random Forest, Support Vector Machine, and Artificial Neural Network. To validate these models, four indicators or values of merit were used: accuracy (acc), precision (p), recall (r), and Matthew's correlation coefficient (mcc). This paper describes dataset preparation and the refinement of the threshold values used for the predictive variable most influential on the class (the copper recovery). Results show both a precision over 98.50% and also the model with the best behavior between the predicted and the real. Finally, the models obtained show the following mean values: acc=94.32, p=88.47, r=99.59, and mcc=2.31. These values are highly competitive as compared with those obtained in similar studies using other approaches in the context.

Keywords: Data analysis; Artificial Intelligence; Machine Learning; Knowledge Engineering; Computers and information processing, Data analysis; Data Processing.

1. Introduction

At present, the copper mining industry in Chile and the world is undergoing big changes owing to the use of modern techniques such as predictive models based on Artificial Intelligence [1, 2, 3]. An example of them is the use of data mining methods to study or predict copper recovery or mineral prospective [4]. These methods relate to data-driving techniques used in several engineering areas, showing good results in most cases [5].

Techniques such as machine-learning algorithms have been recently used in the copper production industry to, for instance, reduce the cost of leaching methods aiming to improve both processes and results [6, 7]. A proper amount of data related to processes and results are necessary for these algorithms to work, but not all machine-learning algorithms work properly in domains as the copper production industry, neither provide good results in all domains, nor work on datasets, mainly concerning suitability of data and accuracy rates [8, 9].

In the copper industry, there are well-known methods used for producing metallic copper via hydrometallurgy: dynamic heap leaching, solvent extraction, and electro-winning. The ultimate goal of these processes is to produce the most amount of copper, saving resources and being the least aggressive for the environment.

 doi:10.20944/preprints202102.0326.v1

Recent studies such as [10] report predictive copper recovery models with 95% accuracy, utilizing artificial neural networks and parameters widely used in industry, such as "Monoclass granulometry", "Irrigation rates", "Total acid added", "Pile high", "Total copper grade", "CO3 grade", "Leaching ratio", "Operation day", and "Soluble copper grade stacked".

Recently, a big number of studies on data-driven approaches have been published. They deal with the use of predictive model techniques for metallic copper recovery processes. For example, [3] report artificial intelligence methods used for developing predictive copper recovery models. For this kind of works, Gradient Boosted Trees (GBT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) being some of the most frequently used. Research on predictive copper models or copper recovery models considers this issue from different perspectives, recent works report models using different data mining techniques [1, 4, 10, 11].

Some studies comparing the results of predictive models used in mining are found in the literature. For example, [12] compare ANN, wavelet neural network (WNN), and SVM to classify the mapping of potential copper points in an Iranian copper mine, the best result being obtained with ANN. Other recent studies compare predictive model performance in other industries [13]. For example, [14] compare the use of ANN, SVM, and RF in the formation of geological reservoirs, while [15] compare ANN and RF in rock drilling and blasting in a mining company. In this study, the authors conclude that the ANN-based model shows the best performance.

In [10] results from the generation of predictive models for copper recovery by leaching is reported. In this study a Lineal model, a Quadratic model, a Cubic model, and an ANN model were developed and the results of the prediction for all of them were compared. The best predictive values were obtained with ANN (97% model accuracy).

More recently, in [3] results from the generation of predictive models for copper recovery by leaching, using a different dataset respect to [10] is reported. In this study, a predictive model for copper recovery using RF was developed and validated with data collected at the plant and in a controlled environment at the lab. The resulting model shows the relevance of the predictive variables over the class (copper recovery), showing 95% accuracy.

A comparison between the models already developed must be made to evaluate the data mining methods most appropriate and generate another model. The literature reports comparisons between models developed with Bayesian networks and ANN, for example: [9, 14], but in other industrial contexts. However, the literature does not show studies on this comparison for copper extraction by leaching. This study aims to compare the performance of three supervised machine learning algorithms used to develop predictive models of copper recovery by leaching. Here, an assessment using cross-validation accuracy among three algorithms is presented: Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest (RF). It aims to compare the results and determine the one contributing with the most knowledge to the leaching process; what can be learned, discover, and predict with each model; and determine the quality measures of each model to compare their performance, all this follow previous works as [3, 16].

The document is organized as follows: Section 2 describes the 1-year copper leaching process studied, by comparing three models to predict copper heap leaching results. Section 2 also compares traditional and machine-learning methods to predict copper recovery by leaching. Section 3 describes artificial intelligence methods used for developing predictive models, particularly, machine learning; the methodology used and its stages; and the experimental and design of the models devised. Section 4 describes the implementation of the models and their validation using performance measures and cross-validation. Section 5 deals with the analysis and discussion of results, along with future research lines. Finally, acknowledgments and bibliography are included.

2. Materials and Methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

As described in [10], SCM Franke mining company uses three processes well-known in the copper mining industry, one of them being of interest in this study, that is, leaching in dynamic cells. In heap leach piles, high mineral variability and irrigation conditions used in the leaching process increases the system entropy, a condition that causes excess uncertainty (close to ±20 [%]) with respects to the copper recovery resulting at the end of the operational cycle.

This usually involves poor planning of the irrigation module start-up, agglomerate and irrigation conditions, drainage distribution, inventories of poor solutions, PLS flow to SX plant, and copper deposited and extracted at the electrowinning (EW) plant, the leaching process being similar in most mining companies. Principal steps on the mining leaching process are represented in Figure 1. The ore is obtained from the open pit mine by extracting copper oxides mostly composed of atacamite. Then, it is sent to the primary crushing section by means of belts.

The ore mixture obtained from the primary crushing is classified according to granulometry and then, sent to either a stockpile, or a secondary or tertiary crusher. If the material meets the granulometry of the company, it is sent to stockpile 1 (defined as coarse stockpile) and later to the secondary crusher.

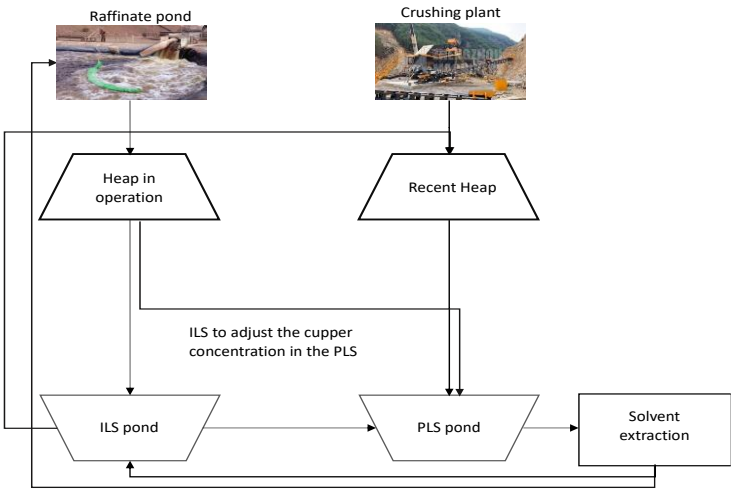

Figure 1. Main steps in the leaching process.

The grain classification is checked again, and the ore is sent to the tertiary crusher, if necessary. Finally, it goes to stockpile 2 (fines), where it is stored to finally enter the agglomeration process. This process takes place in two agglomeration tanks with the same dimensions. Here, the ore gets in contact with a sulfuric acid solution and water. The objective of this process is to form uniform agglomerates of a proper size with the fine material. In addition, it contributes to copper sulfation due to its contact with acid solutions.

At the end of this process, the material is sent to the dry area. Here, stock-piling ends. A radial stockpiler forms a dynamic stockpile used for leaching. It has a maximum of 12 slots, each of them formed by eight modules. Each module has certain dimensions and operates with 10.000-15.000 [t] of agglomerate.

The process to produce copper ends with the drainage of the solution going to PLS. The copper in the PLS solution has an average concentration which is assessed at different times during the stockpile operation.

These measurements allow determining the quality of the leached copper and also help to validate both the planning process of the leaching stockpile life cycle and also how well each part of the process is designed for the stockpile to operate. At this moment, two partial results are obtained: PLS and ILS solutions. The PLS solution is used in the solvent extraction stage, while the ILS solution is used for leaching new modules. Stockpile exploitation is conducted during a period of time called exploitation cycle. These may last 30-45 days in which the tasks described above take place. The goal in these processes is to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

achieve the highest copper production by saving resources with the lowest possible environmental impact.

In this regard, during the leaching stockpile exploitation, the company collected copper recovery data and also data in a controlled environment with stockpile material, although piled in columns and processed at the lab. In this study and similar to [3, 10], the dataset directly obtained from the stockpile is called operational data, while those obtained from the columns at the lab are called column data. Further details about the datasets, how they were obtained, and the pre-process to be used in the models are described in the next section.

2.1. Traditional and Machine Learning processes to predict Copper Recovery by Leaching

One of the most common difficulties companies face when using the leaching technique to recover copper is the lack of historical data for planning the stockpile exploitation and forecasting or planning the operation [3]. Some of the simplest prediction techniques consists in generating or using random variables utilizing iterative simulation and then, conduct goodness-of-fit tests [6].

However, these techniques usually result in high interdependence indexes, possible errors, and subjective interpretation of significance [17] and, therefore, other statistical and mathematical techniques such as time series regression models are used [18]. This kind of techniques have been used in copper mining industry for several decades, but the optimization of results at the domain is possible using machine learning techniques as show previous works as [3, 10, 15, 18].

For example, [19] developed a detailed and accurate simulation model capturing factors that affect air due to copper production and the handling of materials used in the productive activity of an Australian mine. This study was later used in [20] to develop a certain type of statistical model to support the decision to close a mine owing to air quality control.

An example of the use of mathematical and statistical techniques to predict copper recovery by leaching is given in [10], where the authors use Minitab ® software to create three statistical models: Linear, quadratic, and cubic fit. A regression of the best subsets was conducted for each model, obtaining an R2= 92.3% fit for the cubic fit model.On the other hand, machine learning methods are used in studies such as those above, which are further described in the section below.

Machine learning algorithms

Furthermore, there are many cases and discussions on how supervised learning methods (e.g., single-layer neural networks) and deep learning methods (e.g., convolutional neural networks) may be used in different kind of analysis [21, 22] and using different structures of data [23, 24]. Next, a brief description of the supervised machine learning algorithms used in this work is made.

2.2. Artificial Neural Network

Artificial Neural Networks (ANNs) imitate the biological neural network function of the human brain. They can approach every continuous function showing structural formation and complexity. In petrophysics, they are widely used for calculating petrophysical properties, generating synthetic records, and classifying rock facies [25].

The basic elements of a neural network include layers, neurons, and activation functions. The neural network structure is simplified by using a feedback-based neural network with a hidden layer.

The neurons in two adjacent layers are connected by a weighted linear combination and an activation function to introduce non-linearity. The most common activation functions are sigmoid, ReLU, and TanH.

Owing to the dataset size herein, the activation function is limited to a sigmoid function for the hidden layer and a SoftMax function for the output layer. However, ReLU and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

TanH are preferred for deep neural networks because they do not have the vanishing gradient problem [26]. The value of each hidden layer is calculated as:

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i\right) \quad (1)$$

Where x_i are input records, ω_i is the corresponding weight to x_i, and σ is the sigmoid function defined as:

$$\sigma(x) = {}^{1}/_{1+e^{-x}} \quad (2)$$

For each x_i, the cost and gradient functions are calculated using backpropagation and the weights are optimized with a conjugated gradient solver.

2.3. Support Vector Machine

Support Vector Machine (SVM) are classifiers that define a decision limit optimally separating two classes. In linear cases, the decision limit becomes a hyperplane. The construction of the hyperplane separating two classes is based on a sample subset near the decision limit (support vectors). In the linear case, two classes are defined by limits, that is, $w \cdot x\text{-}b \geq 1$ and $w \cdot x\text{-}b \leq -1$. A margin is defined as the distance between these two hyperplanes, $2 / |w|$.

SVM training maximizes the margin between the training points of two different classes. A restriction-delimiting parameter C is used to control penalization when training instances are classified incorrectly. For a high value of C, SVMs tend to generate a smaller margin at the risk of overfit. A small value of C results in more erroneous classifications at the expense of training precision [2, 12].

In [14] the authors describe the use of SVM regression to estimate TOC from string records and compare it with the traditional method ΔlogR. An SVM advantage is that a kernel trick is easy to use for mapping low-dimension data and linear decision limits in a high dimensional space to solve non-linear classification problems. Some of the indicators most used are [27]:

$$K(x, x') = exp\ (x^T x' + c)^d \quad (3)$$

Gaussian radial basis function (RBF) is defined as show in (4). For the RBF and to reduce the number of hyperparameters, only Gauss kernel functions are considered. The kernel is controlled by kernel scale y. A one-to-one classification is implemented in the process to improve the multiclass classification results because the computing cost does not significantly increase with the number of classes.

$$K(x, x') = exp\ (-y\|x - x\|) \quad (4)$$

2.4. Random Forest

Random Forest (RF) is a supervised learning algorithm derived from decision trees (DT), which is frequently used for developing a predictive model [28, 29]. As described in [3, 28, 30], DT is basically a hierarchical set of nodes (beginning with a root-node), where each node contains a decision based on the comparison between an attribute and a threshold value. DT-based learning starts with the observation of an object represented by the branches of a tree and ends with certain conclusions related to the target value of an OBJECT (represented by tree values).

The RF supervised learning algorithm uses supervised learning methodology over a labelled dataset (training set) to make predictions and produce a model. The resulting model can be later used to classify non-labelled data, along with establishing data parameter and threshold values, working on a training dataset.

The method combines the idea of bagging with the random selection of characteristics so as to construct decision trees with controlled variance [28, 31]. One of RF main benefits as a model is that it can be used for determining the importance of the variables in a regression or classification problem intuitively [31].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

This importance is calculated with a metric, according to the impurity decrease in each node used for data partitioning. In the case of classification, the class determined corresponds to the mode of the classes provided by each tree.

3. Materials and Methods

3.1. Data-driving techniques in the copper industry and the Methodology

Currently, there are some cases in which data-driving algorithms are used by the industry of copper production by leaching to analyze data and provide results, such as models for predicting the quality of the copper recovered [3]. Based on data availability, data-driving techniques such as machine-learning algorithms: ANN, SVM, and RF are implemented as copper recovery classifiers. The definitions of these algorithms and their main characteristics applicable in this context are described below.

On the field of copper recovery by leaching, there are a lot of parameters that have an influence on the resulting recovery; some of them can be controlled but other cannot be directly controlled [3]. This section describes the experimentation parameters and the sizes of the used datasets. The methodology used in this study is detailed graphically in Figure 2. Each stage of the methodology is described below:

1. Data preparing. For this first step data from copper heap (called plant) and data from laboratory (called lab) were used, both data sets were acquired in accordance with an experimental design, described below. The datasets contain the features of both process (the leaching process on plant and the leaching process on lab). In both cases, data were prepared according to the results of the predictive variable measurements on heap and lab. In this stage, .csv archives were used with data collected at the plant and the lab. The class 'copper recovery', and associated variables are detailed in Table 1.
2. Labels generation. The second stage consisted in generating labels for the dependent variable Y (copper recovery - see Table 1) in both datasets, according to the threshold values of the other predictive variables. This process was made aiming to prepare the entry data of data mining algorithms. This resulted in datasets with labels for the dependent variable Y.
3. Models generation. The third stage consists in training for developing predictive models, using some parts of the datasets for training and the remaining ones for validation; the details are in Section 4.
4. Validation and analysis of results. The fourth stage deals with the validation and interpretation of results. In this stage, comparative statistical tests using classifier performance measures were utilized to determine the quality of the models developed. The models that resulted from these tasks were analyzed and interpretated, according to expert knowledge and experience in the field of copper mining exploitation by leaching.
5. Prediction. Using the validated models, predictions of copper recovery by leaching were made, and these were compared using the techniques described in the previous step.

The steps of the methodology, from data preparing to validation and analysis of results are summarized in Figure 2.

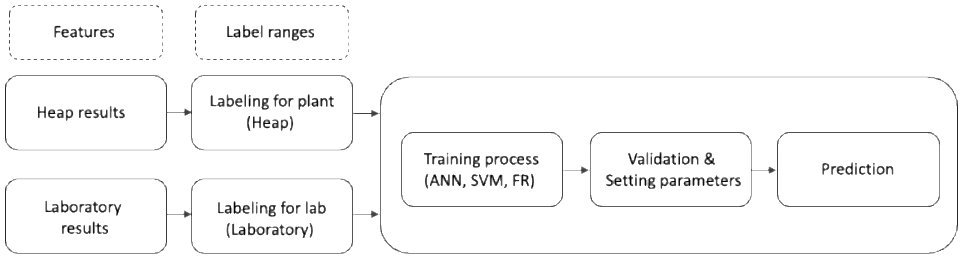

Figure 2. Workflow for Copper recovery prediction using both datasets (Heap and Lab).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

3.2. Data preparation and Experimental Design

For this study the experimental design is based on the previous models described on [3, 10], in the same way, the data used as input for the modes were provided by SCM Franke copper mine, this mine is located in Atacama Region, Chile. More details about this company and their mining processes can be found in [3, 10]. One of the main goals on the leaching process is to obtain the greatest copper production by saving resources and being the least possible aggressive to the environment.

As have been described in [3], the copper leaching process involves tasks as irrigation beginning and maintenance, agglomerate condition evaluation, drainage distribution, pool solution inventory, PLS flow evaluation. Figure 1 illustrate these processes. In later stages the process of distribution and deposition of the material leached at the plant (harvest) are made. At SCM Franke mine the leaching cycles are planned to last 65-100 days.

Two kinds of resulting data related to material leached were considered for this work, the first one is related to a dynamic plant and the second one is related to pilot plant (or piling plant). For each of them, a dataset composed by records used in [10] (2017) and a dataset composed by records used in [3] (2020) were considered.

In order to facilitate the interpretation and relationship with the input data, the following notation was used: A1 and B1 identify the datasets that were collected for the word reported in [10] (2017) and A2 and B2 identify the datasets that were used in the word reported in [3] (2020), the labels A1 and A2 identify the operational data and the labels B1 and B2 identify the laboratory-piling data (pilot data).

Operational data correspond to data that were collected during periods called leaching cycles beginning after soil piling and the starting of the irrigation process, since day 1 to the last day of production. Similarly, laboratory data correspond to production who was conducted in two agglomerate tanks of the same dimensions with a material whose granulometry was less than 13 mm in diameter (more details in [3]).

In this work a total of 20,000 records from operational plant and 15,000 records from piling were analyzed. Using these data, the following datasets were prepared: 4,916 operational records corresponding to 30 leaching cycles during an average of 67 days were cleaned and prepared, and they are **A1**, 3,772 operational records corresponding to 33 leaching cycles during an average of 61 days were cleaned and prepared, and they are **A2**. Also, 3,863 piling records, corresponding to 65 leaching cycles during an average of 61 days were cleaned and prepared, and they are **B1**, and finally 3,030 piling records, corresponding to 63 leaching cycles during an average of 63 days were cleaned and prepared, and they are **B2**.

To develop the predictive models the variables shown in table 1 were used. Each algorithm has its own characteristics and procedure. Therefore, the predictive variables had to be adjusted. Table 1 shows the initial discretization of the variables needed to use the algorithms.

Table 1. Entry parameters to the statistical model.

Var	Description	Low	Normal	High
X1	mono class granulometry (mm)	[11.5-12)	[12-13)	[13-15)
X2	irrigation rates (lhr/m^2)	[6-8)	[8-11)	[11-14)
X3	total acid added (g/l)	[0.5-20)	[20-50)	[50-75)
X4	stocked high (m)	[1-3)	[3-4)	[4-5)
X5	total copper grade (%)	[0.4-0.7)	[0.7-1.2)	[1.2-2.1)
X6	CO3 grade (%)	[0.5-4)	[4-6)	[6-10)
X7	leaching ratio (m^3/TMS)	[0.012-5)	[5-10)	[10-15)
X8	operation (days)	[1-50)	[50-100)	[100-168]
X9	soluble copper grade stacked (%)	[55-65)	[65-80)	[80-98]
X10	Number of stockpiles	-	-	-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

Y	Copper recovery	(%)	[10-65)	[65-80)	(80-98]

To generate the models, the free educational version of Rapidminer 9.7® was used. This software tool allows developing models with desired characteristics and available datasets.

Before to use the algorithm RF, labels for variable Y from the continuous values described on Table 1 were created. Initially three ranges for Y were used, these ranges were principally based on the threshold values of variable X7 indicated on Table 1, but considering the earlies results when the preliminary modes with RF were new umbral values for X7 were found, these three ranges were discarded, and two new ranges were used in accordance with the new threshold values (details are in Section 4.1). These values were: "1" for optimal values of Y, when X7<= 4.2 and, "0" for less-favorable values of Y when X7>4.2. For SVM, the same two ranges for Y were used in order to create two classes. The SVM algorithm takes each dataset and makes a prediction for each entry, about which of the two classes created previously. To meet the above mentioned, it was necessary to make a class separation over the dependent variable (Y). This separation was made considering the characteristics of variable X7.

Variable X7 is a consumption and inventory indicator of the sulfuric acid used in the leaching irrigation process (variable X2) and allows establishing the indicators of the process cost. The relationship between variable X7 and operational costs is as follows: the smaller the variable X7, the smaller the operational costs. However, if variable X7 decreases, it affects copper recovery (dependent variable).

The relationship between variables X and Y is neither bidirectional, nor linear, i.e., an increase in the value of variable X7 does not necessarily result in copper recovery (Y). However, a value greater than 4 does not give evidence of a significant slope in copper recovery increase. But a value 4.2 for variable X7 is considered as a threshold value to separate recovery and generate recovery ranges giving rise to the labels described below: 'low' for a recovery below 45% and X7 in the interval (0.0012- 4.2), and 'high' for a recovery over 45% and X7 in the interval [4.2-15).

In order to argue the previously said, two histograms of dependent variable values (Y) and X7 values are shown in Figure 3. Figure 3 (a) shows examples of copper recovery, while Figure 3 (b) shows examples of variable X7 levels throughout the different leaching cycles. As shown in Figure 3(a), the leaching cycles range from 60 to 100 days, while from day 15 to 35 the slopes of the recovery curves tend to be horizontal. This may involve that the relationship between the irrigation ratio and recovery is not so convenient.

This is particularly observed in the curve exceeding the threshold value of 60% recovery, but with an X7 final value of X7=14.95, which results in a high leaching cycle cost. The above, together with expert knowledge supports the decision to fix value 4.2 as the optimal threshold for X7.

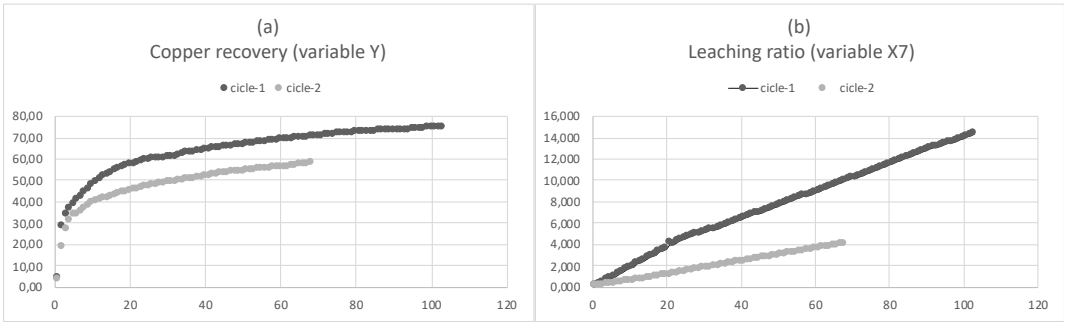

Figure 3. Examples of 2 copper recovery cycles (cicle-1 and cicle-2) and their relationship with variable X7. Part (a) shows levels of leaching ratio for cicle-1 and cicle-2, while part (b) shows the different values of the variable Y (the class) influenced by X7 throughout each leaching cycle.

3.3. Validation using Performance Measures

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

Once the models are developed, they must be validated through different techniques. Often, the models are not validated with the same data used for training the classifier. As reported by [33], the k-fold cross-validation method is a good alternative for validating methods as those developed in this study. In k-fold cross-validation, the (simple) original training set is partitioned into k disjoint sets, one subset being retained for validation and the remaining groups (k-1) randomly selected to be used in training, while the only remaining group of samples is used for creating the cross-validation error function. This function is minimized by changing the values in each training of the automatic learning algorithm [9, 14].

There are several statistical tests using classifier performance measures, which work over the same dataset. Also, there are several alternative performance measures to compare the goodness of classifiers. Here, performance measures that may be applied in the three classification algorithms used were selected [3, 9].

In order to carry out the validation, and based on the data obtained in the experiments, we will obtain a confusion matrix. This matrix will facilitate the analysis needed to determine where classification errors occur. The confusion matrix is a table that shows the distribution of errors in the different categories. The values of merit necessary to evaluate the performance of the classifier to be implemented, will be calculated using this matrix. The confusion matrix is a 2x2 matrix with numerical values a, b, c, and d, which are the result of the classified cases, where a is the sum of the true positive cases, b is the true negatives, c represents the false positives, and d corresponds to the false negatives [22].

Next, the measures of merit of each classifier from previous studies were used in similar way to [3, 9, 10, 30]. The measures of merit used in this study help to determine the quality of the predictive models developed and are based on data from the confusion matrix and the result of training with each classification algorithm. These values of merit are the following:

1. accuracy (acc): corresponding to the ratio of correctly classified examples from all the examples in the dataset [39]. This indicator can be calculated with the confusion matrix data, according to (4), assuming that the dataset is not empty.

$$\text{acc} = (a + d)/(a + b + c + d) \tag{4}$$

2. Precision (p): is the proportion of true positives (a) among the elements predicted as positive. Conceptually, precision refers to the dispersion of the set of values obtained from the repeated measurements of a quantity. Specifically, a high precision value (p) implies a low dispersion in measurements. This indicator can be calculated according to (5), assuming a+b≠0.

$$\text{p} = (a/a + \ b) * 100 \tag{5}$$

3. Recall (r): is the proportion of true positives predicted among all elements classified as positive, that is, the fraction of relevant instances classified. Recall can be calculated according to (6), assuming a+c≠0.

$$\text{r} = (a + d)/(a + c) \tag{6}$$

4. Matthew's correlation coefficient (mcc): is an indicator relating the predicted versus the real, creating a balance between the classes, considering the instances correctly and incorrectly classified into classes quite different in size and with a significant number of observations [35]. The mcc value can be calculated according to (7), assuming that the destination dividend (τ) is not zero.

$$\text{mcc} = \big((a + b) - (c + d)\big)/\tau \tag{7}$$

where

$$\tau = \sqrt{((a + d) * (a + c) * (b + d) * (b + c)} \tag{8}$$

 doi:10.20944/preprints202102.0326.v1

In each of these models, the values of merit of Eq. (4) to (7) were calculated, being useful for comparing the goodness of the models developed. These calculations, interpretations, and comparisons are described in the next section.

4. Results and Discussion

This section describes the results obtained. As mentioned above, in order to facilitate their interpretation and relationship with initial data, the following notation was used: dataset A1 is the stockpile dataset used in [10] and dataset A2 the stockpile dataset used in [3]. Analogously, dataset B1 is the pilot dataset used in [10] and dataset B2 the pilot dataset used in [3].

Table 2 shows the RF total classification and the datasets for each copper recovery label ('low', 'medium', 'high'), corresponding to datasets A1 and A2, and table 3 shows the corresponding information for datasets B1 and B2. In table 2, the results show that classification precision is always over 95%, and on the other hand, table 3 shows that classification precision is over 99% for the 3 ranges. Additionally, for both tables the difference between the real and predicted is lower than 30% for all class classes.

This may be interpreted as a good approximation of the RF-generated models to what is actually observed in each dataset. It also shows that the classification precision is always over 97%, while the real and predicted is below 30% for all the class cases.

Figure 4 shows the real vs. the predicted in each dataset, according to the classification labels. The abovementioned is related to the interpretability of results from the RF-generated models.

Figure 5 shows no significant changes between the values of merit in any of the four scenarios. These data, together with the interpretation of trees described in Figure 6, show that separating the datasets is not significant, concerning the three values of the labels in the class. Rather, considering the threshold value X7=4.1, it is possible to optimize the separation into two groups in the class as follows: 'low' in the class variable for values X7<4.1 and 'high'.

Figure 6 (a) shows a result for dataset B1, clearly illustrating that 'medium' and 'high' groups may be regrouped into one group under threshold X7>=4.097, while Figure 6 (b) shows a result for dataset A2, illustrating that 'medium' and 'high' groups may be regrouped into one group under threshold X7>=4.108.

For RF training, the value of parameter 'tree number' was set at 50, while for parameter 'criterion', the gain ratio value was selected at a maximal depth of 10 for each model.

The resulting values of parameter 'classification error' for each training were as follows: 0.0014 for training with dataset A1; 0.0169 for training with dataset A2; 0.0403 for training with dataset B1; and 0.0715 for training with dataset B2.

The value of parameter 'classification error' increases from training with the first dataset to the last one, but no relationship was found between this result and the possible interpretations of the model results.

For SVM training, the value of C was set at 0.0. For training with ANN and each dataset, an operator called feed-forward neural network from Rapidminer© was used, with training cycles = 200 and a value of learning rate = 0.01. Table 4 shows the values of merit for the classifications of these algorithms. These values were obtained from the classification values in the confusion matrixes shown in tables 5-7.

4.1. Discussion

In this study, a total of twelve predictive modes were generated, in detail: four predictive models using RF, four with SVM and four using ANN. For these optimizations 80% data were used for cross-validation and 20% for validation, in similar way to [3]. Added, four measures of merit including acc, p, r and mcc were used or each algorithm to evaluate the quality of all these machine learning methods (RF, SVM and ANN).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

Table 2. Confusion matrix; accuracy and classification error values for datasets A1 and A2.

	Dataset A1 accuracy: 99.97% +/- 0.09% classification error: 0.03% +/- 0.09% (micro average: 0.05%)				Dataset A2 accuracy: 99.94% +/- 0.12% classification error: 0.07% +/- 0.11% (micro average: 0.06%)			
	true low	true medium	true high	class precision	true low	true medium	true high	class precision
pred. low	2896	0	0	100.00%	2893	0	0	100.00%
pred. medium	1	567	0	99.82%	0	371	0	100.00%
pred. high	0	1	19	95.00%	0	1	175	99.43%
class recall	99.97%	99.82%	100.00%		100.00%	99.73%	100.00%	

Table 3. Confusion matrix; accuracy and classification error values for datasets B1 and B2

	Dataset B1 accuracy: 99.93% +/- 0.16% classification error: 0.07% +/- 0.16% (micro average: 0.07%)				**Dataset B2** accuracy: 99.93% +/- 0.14% classification error: 0.07% +/- 0.14% (micro average: 0.07%)			
	true low	true medium	true high	class precision	true low	true medium	true high	class precision
pred. low	1734	1	0	99.94%	1720	0	0	100.00%
pred. medium	1	511	0	99.80%	0	439	1	99.77%
pred. high	0	0	456	100.00%	0	1	870	99.89%
class recall	99.94%	99.80%	100.00%		100.00%	99.77%	99.89%	

First, using the datasets as input, RF algorithm and three labels for the variable Y, predictive models for Y were generated. Tables 2 shows the classification results using A1 and A2 with 99.95% as final accuracy, similar results using B1 and B2 are show in Table 3, where the final accuracy is 99.93%. For both the records labeled as "medium" and "low" for X7 can be regrouped in "low" because not exceeding the threshold value 4.2 for values of Y in the range 60-80% of recovery.

Meanwhile, figure 4 shows the real vs. the predicted in each dataset, according to the classification labels, this information together with the interpretation of trees described in Figure 5, show that separating the datasets is not significant, concerning the three values of the labels in the class, because Figure 5(a) shows that the limited value between "high" and "medium" is 4.097 and part (b) shows that the limited value is 4.1. Rather, considering the threshold value X7=4.2, it is possible to optimize the separation into two groups in the class as follows: "low" in the class variable for values X7<4.2 and "high" for all other cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

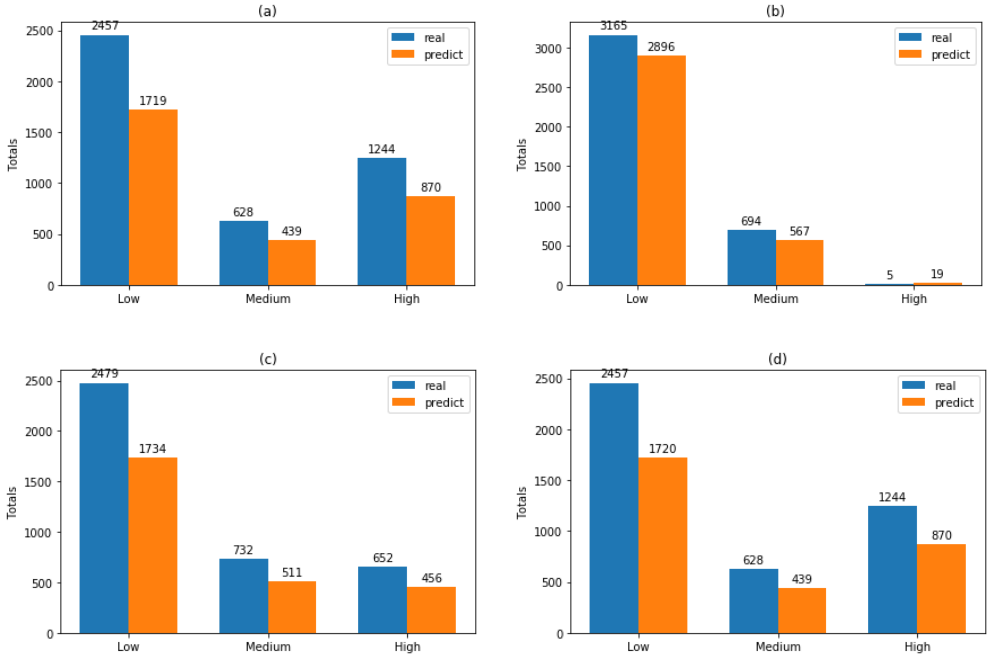

Figure 4. Real vs. predicted values with RF classifier for each dataset, using labels "low", "medium", and "high" for the dependent variables.

Figure 5. Two tree structures developed with RF, using two different datasets. Figure 5(a) was generated using as input the dataset A1 and side (B) was generated with the dataset B2.

Table 4 shows the values of merit for validating the quality of the model classification results. This table shows the classification precision (class_precision) in all models is over 98.50%, while an ideal classification precision is 100%. The mean value of acc = 0.943 indicating that almost all the samples in the datasets were correctly classified, while the worst absolute value in all the models was for dataset B1.

The partial result shows in table 4, related to dataset B1, might be indicating that the data source must be checked to identify its cause. The low value of the variable of merit associated with each case of acc for each dataset may be due to high dispersion of the samples in dataset B1. The low value of p in the results of dataset B2 may be associated with high data dispersion in the data from the pilot dataset.

Table 4. Values of merit for predictive models developed with algorithms FR, SVM, and ANN.

	Dataset	class_ precision	acc	p	r	mcc
RF	A1	99.700	0.9491	99.7558	1.0000	0.0563

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

SVM	A2	100.000	0.9945	99.4548	1.0000	0.0529
	B1	100.000	**0.8214**	**83.1261**	1.0000	0.0026
	B2	99.850	0.9993	**71.3012**	1.0000	**0.0016**
	A1	98.635	0.9491	99.7558	0.9955	0.0571
	A2	98.717	0.9945	99.4548	0.9878	0.0537
	B1	98.607	**0.8314**	**83.1361**	0.9877	0.0025
	B2	99.381	0.9993	**71.3712**	0.9878	0.0015
ANN	A1	99.381	0.9491	99.8350	1.0000	0.0042
	A2	99.921	0.9946	99.6177	0.9984	0.0437
	B1	99.381	**0.8312**	**83.5156**	0.9954	**0.0018**
	B2	99.381	0.9958	**71.4451**	0.9981	**0.0011**

For table 4, the result of the variable of merit mcc is also remarkable. Since all the values are low, the model prediction is quite close to real. Considering this result, it is then possible to interpret that the datasets and the preparation of the predictive variables and labels are correct.

As is known RF, SVM and ANN are probabilistic models that provide fa powerful formalism for representation, reasoning and learning under uncertainty [36]. Considering the results given in table 4, ANNs show the best capability to reasoning over the datasets, since the best mcc values correspond to these models, while mcc values for RF and SVM models are quite similar and close, this means that RF and SVM have similar behaviors over the data used in this work.

Notably, the worst values for acc were with dataset B1, while the worst values for p were with dataset B2, in relation to the parameter r, no significant differences have been observed among the models. On the other hand, the well-classified figures (in the models' confusion matrix tables) are above 95.98% as shown in tables 5, 6 and 7.

Table 5 shows the classification obtained with RF for all datasets with labels “high” and “low” for the class Y. In particular it should be noted that the final accuracy is over 99,55%. The best classification obtained has been for the label “low”, with a total of 100% of precision for the 99,97% of the total of the cases for this classification.

Similarly, Table 6 shows the classification obtained with SVM. The final accuracy is over 98,65% this value is lower than the value obtained by using FR. In this case the best classification obtained has been for the label “low”, with a total of 98,69% of precision for the 99,12% of the total of the cases for this classification.

In the same way, Table 7 shows the classification obtained with ANN. For this case the final accuracy is over 99,51% this value is also above that the value obtained by using SVM. Similarly, the best classification obtained has been for the label “low”, with a total of 99,41% of precision for the 99,57% of the total of the cases for this classification.

Table 5. RF Confusion matrix for datasets A1, A2, B1, and B2.

	Dataset A1			Dataset A2		
	true low	true high	class precision	true low	true high	class precision
pred. low	3265	0	100.00%	3466	0	100.00%
pred. high	1	275	99.43%	0	1023	100.00%
class recall	99.97%	100.00%		100.00%	100.00%	
	Dataset B1			**Dataset B2**		
	true low	true high	class precision	true low	true high	class precision
pred. low	2248	0	100.00%	2159	1	99.95%
pred. high	0	456	100.00%	1	869	99.89%
class recall	100.00%	100.00%		99.95%	99.89%	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

Table 6. SVM Confusion matrix for datasets A1, A2, B1, and B2.

	Dataset A1			Dataset A2		
	true low	true high	class precision	true low	true high	class precision
pred. low	3660	26	99.45%	3074	50	98.40%
pred. high	5	224	97.82%	12	894	99.00%
class recall	99.89%	90.60%		99.61%	95.98%	
	Dataset B1			**Dataset B2**		
	true low	true high	class precision	true low	true high	class precision
pred. low	3203	48	98.52%	2074	50	98.40%
pred. high	8	604	98.69%	12	894	99.00%
class recall	99.75%	92.64%		99.61%	95.98%	

Table 7. ANN Confusion matrix for datasets A1, A2, B1, and B2.

	Dataset A1			**Dataset A2**		
	true low	true high	class precision	true low	true high	class precision
pred. low	3662	2	99.96%	4951	2	99.84%
pred. high	3	248	98.80%	0	1092	100.00%
class recall	99.94%	99.20%		100%	98.42%	
	Dataset B1			**Dataset B2**		
	true low	true high	class precision	true low	true high	class precision
pred. low	3207	19	99.41%	2771	9	99.61%
pred. high	4	633	99.37%	16	1231	99.51%
class recall	99.88%	97.09%		99.81%	99.03%	

All these tables (5, 6 and 7) are showing that the classification precision for all generated models is over 90.00%, while the gap between the real and predicted is below 10% for all the class cases.

An additional prelaminar conclusion based on the analysis of the obtained merit values seem to be mostly balanced, without classes bringing down the performance in a significative way. This is a very important feature for models when imbalanced data is used as input, as indicate in [37].

In general, and as a result of the above analysis, it is possible to conclude that the classifications obtained when SVM and ANN were used are very closer to the real number of records of the datasets, but the classifications obtained using the RF algorithm are more exactly because the predicted values are very close to 100%. Figure 6 illustrates the contrast describe above.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 February 2021 doi:10.20944/preprints202102.0326.v1

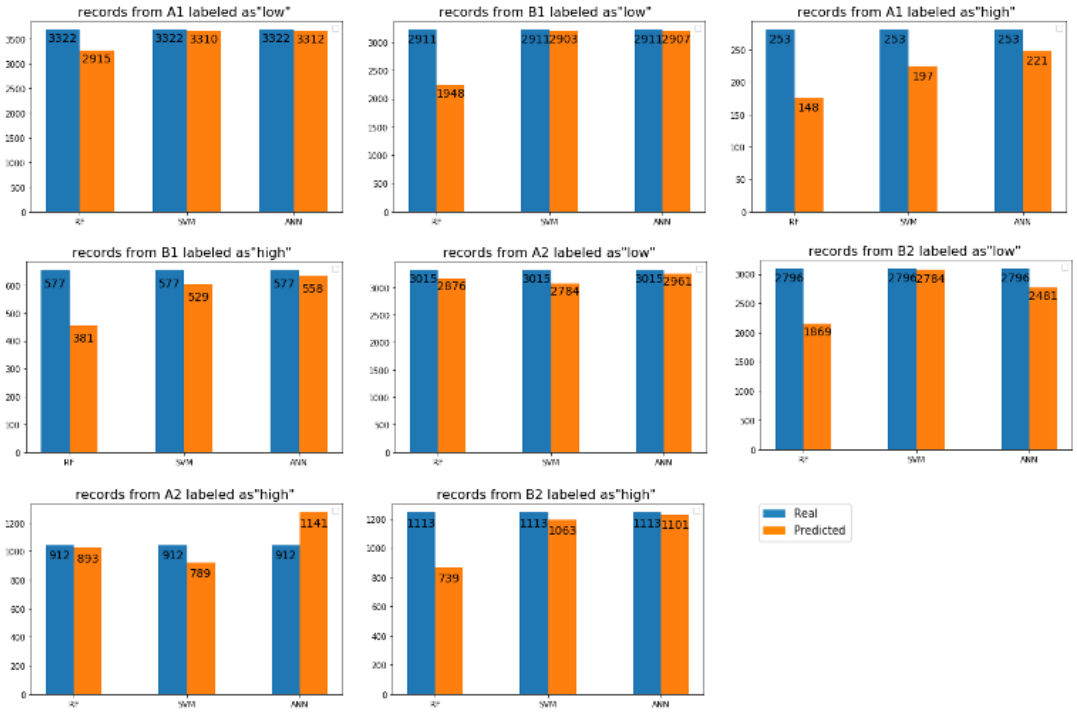

Figure 6. Comparative of real vs predicted values for each dataset and considering the labels 'low' and 'high' for each dataset.

Conclusions

This study compares three supervised machine learning algorithms for classifying copper recovery quality prediction in a leaching process, using real data collected in a copper mine of the north of Chile. In summary, the three artificial intelligence methods – RF, SVM, and ANN – may be used to develop predictive models for copper recovery, identifying and validating the most influential predictive variables of class Y (copper recovery). This study uses datasets prepared and tested in previous studies to develop predictive models detailed in [3, 10]. One of the main contributions of this study is the use of four datasets: two related to real operational data and two related to laboratory-piling (pilot), used by developing predictive models with RF, SVM, and ANN.

Resulting models were validated using the cross-validation method, utilized in training an 80-20 k-folder cross-validation. This allowed developing validated and reliable models with a minimal error value in trainings. This resulted in an average accuracy over 95% in all the predictive copper recovery models, using the four datasets prepared for this study, and an average precision over 98% in all the classifier trainings. Also, measures of merit (acc, p, r, mcc) for each classifier were used, in order to determine the quality of the predictive models developed. The models obtained show the following mean values: acc=94.32, p=88.47, r=99.59, and mcc=2.31. These values are highly competitive as compared with those obtained in similar studies using other approaches in the context.

For this work some interesting results respect to RF, SVM, and ANN have been achieved, for example, the parameter optimization with SVM models is noticeable for handling no-linear data, and this method was useful to draw decision boundaries between data points from different classes and separate them with maximum margin. RF method was also useful for identifying the threshold values of the predictive variables that influence class values in the prediction. Another interesting conclusion is related to the closeness of predictions to real results. Particularly in this case, the best method was ANN, confirming its potentiality for developing machine learning models accurate for copper mining exploitation. The above is not detrimental to the results of the other methods used (RF and SVM) since the results obtained with the last two are also quite good and accurate.

Filially, as has been said throughout this document, the use of artificial intelligence methods, in particular of three machine learning algorithms in conjunction with real data is not very common in the copper industry. Thus, we believe the comparative study among these machine learning methods builds an excellent platform for future studies in this area.

Author Contributions: C.L. was responsible for the data collected and the data cleaning. C.L and V.F were responsible to the selection and the validation of parameters used on the models generation, according to the chemical process of copper leaching. V.F. was responsible to the pre-processing, the training whit the Artificial Intelligence models, the validation and interpretation of results. All authors contributed to the design of the experimental design and writing and reviewing the final manuscript All the authors have read and agreed with the final version and published of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank the SCM Franke company for the material support provided and their collab-oration to this study, particularly their support for obtaining and storing data from the copper production process using the leaching method, as well as data from the pilot plant (column simulation) and are thankful for the financial support provided by Universidad Católica del Norte.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abedi, M.; Norouzi, G. H.; Bahroudi A. Support vector machine for multi–classification of mineral prospectivity areas. Computers & Geosciences 2012, 46, 272–283.
2. Abedi, M.; Torabi, S.A; Norouzi, G. H.; Hamzeh M.; Elyasi G. R. PROMETHEE II: A knowledge–driven method for copper exploration, Computers & Geosciences 2012, 46, 255–263.
3. Flores, V.; Keith B; Leiva, C. Using Artificial Intelligence Techniques to Improve the Prediction of Copper Recovery by Leaching, Journal of Sensors 2020, 1–6.
4. Sun, T.; Chen, F.; Zhong, I.; Liu, W.; Wang, Y. GIS–based mineral prospectively mapping using machine learning methods: A case study from Tongling ore district, eastern China, Ore Geology Reviews 2019, 109, 26–49.
5. Peters, T. Data–driven science and engineering: machine learning, dynamical systems, and control, Cambridge University Press, Cambridge, England 2019.
6. Barga, R.; Fontama V.; Tok W. H. Predictive analytics with Microsoft Azure machine learning, Apress, Berkeley, CA., USA, 2015, 221-241.
7. Milivojevic, M.; Stopic, S.; Friedrich, B.; Drndarevic, D. Computer modeling of high–pressure leaching process of nickel laterite by design of experiments and neural networks, Int. Journal of Minerals, Metallurgy, and Materials 2012, 19(7), 584–594.
8. Çığşar, B.; Ünal, D. Comparison of data mining classification algorithms determining the default risk, *Scientific Programming* 2019, 1–12.
9. Flores, V.; Correa, M. Performance of Predicting Surface Quality Model Using Softcomputing, a Comparative Study of Results, In Proceeding of the International Work–Conference on the Interplay Between Natural and Artificial Computation, Spain, 233–242, 2017.
10. Leiva, C.; Flores, V.; Salgado, F.; Poblete D.; Acuña, A. Applying Softcomputing for copper recovery in leaching process, *Scientific Programming* 2017, 1–11.
11. Tan, M.; Song, X.; Yang, X.; Wu, Q. Support–Vector–Regression Machine Technology for Total Organic Carbon Content Prediction from Wireline Logs in Organic Shale: A Comparative Study, *Journal of Natural Gas Science and Engineering* 2015, 26, 792–802.
12. Saljoughi, B.S.; Hezarkhani A. A comparative analysis of artificial neural network (ANN), wavelet neural network (WNN), and support vector machine (SVM) data–driven models to mineral potential mapping for copper mineralizations in the Shahr–e–Babak region, Kerman, Iran, *Applied Geomatics*2018, 10(3), 229–256.
13. Piccarozzi, M., Aquilani, B. and Gatti, C. Industry 4.0 in management studies: A systematic literature review. *Sustainability* 2018, 10(10), 381–3821.
14. Deng, T.; Xu, C.; Jobe, D.; Xu, R. Comparative Study of Three Supervised Machine–Learning Algorithms for Classifying Carbonate Vuggy Facies in the Kansas Arbuckle Formation, *Petrophysics* 2019, 60(6), 838–853.
15. Sadeghi, F.; Monjezi M.; Armaghani, D. Evaluation and optimization of prediction of toe that arises from mine blasting operation using various soft computing techniques, *Natural Resources Research* 2020, 29(2), 887–903.
16. Zadeh, L. A. Fuzzy logic, *Neural Networks and soft computing*, Communications of the ACM 1984, 37(3), 77–84.

17. Grove, W.M.; Meehl, P.E. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy, *Psychology, Public Policy, and Law* 1996, 2(2), p. 293.
18. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms, *Adv. in Neural Information Processing Systems* 2012, 25, 2951–2959.
19. Welgama, P.; Mills, R.G.; Aboura, K.; Struthers, A.; Tucker, D. Evaluation of Options to Improve Copper Production, In Proceeding of the 6th International Conference on Manufacturing Engineering, Melbourne, Australia, 219–223, 1995.
20. Aboura, K. A Statistical Model for Shutdowns due to Air Quality Control for a Copper Production Decision Support System, *Organizacija* 2015, 48, 198–202.
21. Mahmoud, A.; Elkatatny, S.; Mahmoud, M.; Abouelresh, M.; Ali, A. Determination of the Total Organic Carbon (TOC) Based on Conventional Well Logs Using Artificial Neural Network, *Int. Journal of Coal Geology* 2017, 179, 72–80, 2017.
22. Xu, C.; Misra, S.; Srinivasan, P.; Ma, S. When Petrophysics Meets Big Data: What Can Machine Do? In Proceeding of the SPE Middle East Oil and Gas Show and Conference. Society of Petroleum Engineers, Manama, Bahrain, 18–21, 2019.
23. He, H; Bai, Y.; Garcia E.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, In Proceeding of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong 2008, 1322–1328.
24. Koul, N. Learning predictive models from massive, semantically disparate data, Ph.D. dissertation, Iowa State University, Iowa, USA, 2011.
25. Saneifar, M.; Aranibar, A.; Heidari, Z. Rock Classification in the Haynesville Shale Based on Petrophysical and Elastic Properties Estimated from Well Logs, *Interpretation* 2015, 3, 65–75.
26. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifer Neural Networks, In Proceedings of the fourteenth international conference on artificial intelligence and statistics, FL, USA, 315–323 April 2011.
27. Trevor, H.; Tibshirani, R.; Friedman, J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer New York 2003.
28. Breiman, I; Random forests, *Machine Learning* 2001, 45(1), 5–32.
29. Lior, R. Data mining with decision trees: theory and applications, *World scientific* 2014, 81, 11–50.
30. Ho, T.K. Random decision forests, In Proceeding of the 3rd international conference on document analysis and recognition, Australia 1995, 278–282.
31. Barandiaran, I. The random subspace method for constructing decision forests, IEEE transactions on pattern analysis and machine intelligence 1998, 20(8), 832–844.
32. Hofmann, H.; Klinkenberg R. RapidMiner: DataMining Use Cases and Business Analytics Applications, CRC Pres Teylor & Francys Group, FL USA, 2013.
33. Arlot, S.; CelisseA. A Survey of Cross–Validation Procedures for Model Selection, *Statistics Surveys* 2010, 4, 40–79.
34. Chow, C.K.; Liu, C. Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory* 1968, 14(3), 462–467.
35. Bisgin, H.; Bera, T.; Ding, H.; Semey, H.G.; Wu, L.; Liu, L.; Tong W. Comparing SVM and ANN based machine learning methods for species identification of food contaminating beetles, *Scientific reports* 2018, 8(1), 1–12.
36. Yan, B.; Xu, D.; Chen, T.; Wang, M. Leachability characteristic of heavy metals and associated health risk study in typical copper mining–impacted sediments, *Chemosphere* 2020, 239, Article 124773.
37. Xu, C.; Misra, S.; Srinivasan, P.; Ma, S. When Petrophysics Meets Big Data: What Can Machine Do? In Proceeding of the SPE Middle East Oil and Gas Show and Conference. Society of Petroleum Engineers, Manama, Bahrain, 18–21, 2019.