Supervised Training and Unsupervised Machine Learning Testing for Cobalt Dissolution from Complex Cu-Co Ores in Acid Lixiviants

Onesimus R. Baloyi[1], Prof. Antoine F. Mulaba- Bafubiandi[1,2]

***Abstract*—** The ever-increasing application of machine learning and artificial intelligence techniques in research has had a pivotal impact on the existing body of knowledge. In the presence of huge data set available, application of machine learning probes a better handling of the data. As such, this paper discusses the philosophies involved within machine learning. The paper the discusses the supervised training and unsupervised testing stages. The study focused on a supervised model (logistic linear regression) to perform a classification predictive analysis on cobalt dissolution. It further detailed the stages of data preparation and handling of the data upon developing the model. The model's performance was validated by using performance metrics such as accuracy, precision, and recall. The model performed good obtaining an accuracy value of 75%, the precision for the "Low Yield" class is 45,45%. The recall for the "Low Yield" class is 62,5%. The precision for the "High Yield" class is 88%. The recall for the "High Yield" class is approximately 78,57%. The model showed to perform well when predicting high yield instances as compared to low yield instances. As a result, the values obtained showed that the model requires some improvements but its implementation towards the dissolution of cobalt proved to be significant as it can describe high or low yield outcomes based on specific dissolution conditions.

***Keywords*—** cobalt, data, dissolution, machine-learning, models, supervised, and training.

I. Introduction

MACHINE LEARNING over the years has seen a great deal of application in minerals processing. Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed. The two most widely used techniques, supervised and unsupervised machine learning, have significantly improved decision making in metallurgy. This ranges from the application of Artificial Neural Networks, Principal component analysis etc. The application of Machine learning in hydrometallurgical processes is still an emerging concept and proves to be a useful tool in decision making.

In recent years, Cobalt has become the center of technological developments Owning to its use in energy storage devices and application to technological developments, this notion has spiked the worth and price of cobalt [1], [2]. Cobalt has distinctive properties that are suitable for engine parts for aircrafts, permanent magnets, glasses, and ceramic pigments [3]. As a result, the demand for this commodity has risen over the years, it has become more valuable and sold and more costly [4].

With more developments on electric cars, the demand for cobalt will rise exponentially, moreover cobalt will be essential in the development of lithium-ion batteries, as it plays a major role in the functionality of the battery [2]. As a result, this necessitates more research work on the extraction of this commodity to find more economical and environmentally friendly processing routes or conditions [2]–[4]. The RD-Congolese Copperbelt hosts one of the largest Cu and Co reserves in the world, with varying mineralogy from region to region, frankly in some cases from the same deposit [1]. Furthermore, this results in different hydrometallurgical response during the treatment of these mineralogically different ores.

The extraction of Co from complex Cu-Co ores requires intuitive thought process and decision making. As such, the use of implemented decision-making tools becomes an intuitive approach towards prediction of results, re-using domain/old knowledge results on new cases for optimization purposes, this includes process, time and cost optimization. These challenges faced therefore allow an in-depth study to ensure sustainable management of hydrometallurgical processes. Therefore, the use of Machine learning techniques is inevitable to answer the problems raised.

II. Background

A. The leaching of Cobalt

The major or main cobalt sulphide ore is carrolite ($CuCo_2S_4$) which is processed through flotation and with the cobalt oxide ore, the main sources are asbolane (CoO) ($(Ni,Co)^{2-}\times Mn^{4+}(O,OH)_4nH_2O$), smaltite ($CoAs_2$), and heterogenite (when crystallized forms to stainierite $Co_2O_3.H_2O$ or $CoO._2Co_2O_3.6H_2O$ in an amorphous form) [7]. It is recorded that these ores require specific conditions to achieve an

Onesimus Baloyi[1] is with the Mineral Processing and Technology Research Centre, Department of Metallurgy, School Mining, Metallurgy and Chemical Engineering, Faculty of Engineering and The Built Environment, University of Johannesburg, P.O Box 17011, Doornfontein, 2028, South Africa

Antoine F. Mulaba- Bafubiandi[1,2], Mineral Processing and Technology Research Centre, Department of Metallurgy, School Mining, Metallurgy and Chemical Engineering, Faculty of Engineering and The Built Environment, University of Johannesburg, P.O Box 17011, Doornfontein, 2028, South Africa

Faculte de Polytechniques, Universite de Mbuji-Mayi , BP 225, Mbuji-Mayi, Kasai, Republique Democratique du Congo.

optimum dissolution, these require the presence of reducing agents or strong leaching acids. Although the aforementioned ores are within abundance in the Copperbelt deposit, there are other sources of cobalt that are present in the deposit, the likes of malachite ($Cu_2CO_3(OH)_2$), azurite ($Cu_3(CO_3)_2(OH)_2$), and kolwezite ($((Cu,Co)_2(CO_3)(OH)_2$) are easily leached under acidic condition [8]–[9]. It is important to note the mineralogical difference of the Copperbelt deposit, as this results in different processing conditions which then will lead to a varying dataset.

Cobalt in most cases is extracted as a by-product of copper and nickel in nickel containing laterites and nickel-copper sulphide deposits [10]. The extraction of cobalt or copper from oxidized minerals is achieved using different leaching reagents such as sulfuric acid (H_2SO_4), nitric acid (HNO_3), hydrochloric acid (HCl), ammoniacal system, and hydrofluoric acid (HF) or mixtures of these agents [6]. The use of sulfuric acid is prominent, as this is a result of the corrosiveness of the equipment and the selectivity criteria.

As mentioned above, sulphuric acid is widely used as an agent to leach-oxidized cobalt ore. This is denoted by the reaction case below.

$$CoO + H_2SO_4 \rightarrow CoSO_4 + H_2O.$$

Analysis shows that the main mineral of cobalt, heterogenite for this case, contains Co (II) and Co (III). Moreover, the difficulty lies within the treatability of these forms of cobalt, Co (II) is said to be easy to treat whereas Co (II) has some difficulties with regards to its treatment [7]. Therefore, the need to use reducing agents is prevalent to increase the solubility of cobalt [11]. Agents such as ferrous ions, sulphur dioxide, sodium metabisulphite, iron, copper, and more have been used for this case mentioned. The most prevalent reducing agents have been Cu, Fe, $Na_2S_2O_5$, and Fe^{2+} [1]. The reducing agents reacting with sulphuric acid medium result into the four reactions described below,

$$Co_2O_3 + CuO + 3H_2SO_4 \rightarrow 2CoSO_4 + CuSO_4 + 3H_2O$$

$$Co_2O_3 + FeO + 3H_2SO_4 \rightarrow 2CoSO_4 + FeSO_4 + 3H_2O$$

$$2Co_2O_3 + Na_2S_2O_5 + 3H_2SO_4 \rightarrow 4CoSO_4 + Na_2SO_4 + 3H_2O$$

$$Co_2O_3 + 2FeSO_4 + 3H_2SO_4 \rightarrow 2CoSO_4 + Fe_2(SO_4)_3 + 3H_2O$$

These reducing agents yield good results in terms of the solubility of cobalt. The success rate of using these reducing agents is evident but with some of these reagents environmental concerns arise. As a result, alternatives should be considered. This is evident with the study conducted by reference [12] where they investigated the dissolution of Co (III) in the presence of $Na_2S_2O_5$ as a reducing agent. The results of the study were satisfactory although the emission of SO2 was recorded to have an environmental impact, as a result a more environmentally friendly reducing agent was considered. In this perspective, the use of Fe^{2+} instead of $Na_2S_2O_5$ was under investigation. Mbuya et al. exploited the Fe^{2+}-Fe^{3+} loop created in the reaction medium from a composite of minerals. These studies were able to show that it is possible to dissolve Co (III) using Fe^{2+} ions from a Fe-bearing mineral. The conventional leaching of cobalt from ores or concentrates involves the presence of acidic-ferrous sulphate of which various physio-chemical parameters such as pulp density, stirring speed, temperature, pH, reaction time, and ferrous sulphate dosage govern the leaching process [13]–[15].

B. Machine learning

1) Introduction

Machine learning is given a much broader definition owning to its broad application across all disciplines. It is defined as a field of study that gives computers the ability to learn without being explicitly programmed [18]. It is further said to be a computer program that learns from an experience E with respect to some task T and some performance P, if its performance on T, as measured by P, improves with experience E [20].

The fundamental concept of machine learning is to develop models that mimic and generalize data. The understanding of this concept is developed from assigning a computer program to learn from experience when the measurable performance in these tasks improves as it gains more experience in executing tasks [18]. In essence, machine learning tools equip programs with the ability to learn and adapt to different instances [17]. Human error or mistakes is the root cause for bad analysis results and mistakes generated when trying to establish relationships between multiple features, henceforth the development of machine learning techniques has curbed the unfortunate handling of analysis, machine learning is therefore used to train machines to be able to handle data (ranging from small to large data) efficiently [17].

2) Categories of machine learning

Machine learning consists of mainly two broad categories which include supervised and unsupervised machine learning. These categories are applicable for different real-life problems, within them various algorithms are applied to make predictions. Supervised machine learning is tasked by learning from a function that links the input to an output based on sampled input-output pairs [18]. It makes use of the labelled training data set consisting of set datapoints examples. Furthermore, this learning requires external assistance. With unsupervised machine learning, the concept applied here is that there is no correct answer nor the teacher (external assistance). The conceptualization lies within the algorithms to detect trends within the data and presents the interesting structure on the data. the algorithms in this case learn few features in the data [17]. As new data is introduced it makes use of previously learned features from the data to recognize the class to which the data belongs.

C. Review on Supervised machine learning algorithms

Supervised machine learning makes use of two main categories of problems mainly classification and regression. With classification the prediction is done with label or class such as when emails are classified into spam and no-spam mail classes [21], [22]. Moreover, with regression prediction is done using continuous quantities or infinite possibilities like change

in weather conditions. For this research, a few of the algorithms will be discussed since it would take years to investigate all of them. It is to be noted that some have already been studied as such they won't be investigated.

1) Machine learning problem types

In machine learning a type of problem is applied particularly in supervised machine learning. In instance, based on the data type and research objective this therefore dictates which machine learning algorithm is suitable for that problem. In supervised machine learning there are two techniques applied to different data related problems namely, classification and regression. In this section, both techniques will be discussed to see where they are applicable and what type of result, they generate.

2) Classification problem type

Classification is a machine learning technique that is used to forecast group instances, investigates the formation of group memberships [18]. Its use is quite imperative towards industrial use for forecasting future instances, this could include, future production levels, changes to the processes etc. This technique receives much recognition for its ability to predict future occurrences and helps with future planning and knowledge discovery [20].

Although this technique seems to be helpful it has its own limitations such as handling missing data. It is well known that missing data has a huge impact towards creating problems during the training and classification phases. Classification can be used for both structured and unstructured dataset. Members of a dataset are classified according to some given label or category and for new input instances, the class or label that will be assigned to it is predicted by this technique. A classifier algorithm is an algorithm that learns from the training set and then assigns new data points to a particular class [23], [24]. A classification model concludes some valid mapping function from training dataset and predicts the class label with the help of the mapping function for the new data entry. An attribute or feature is a parameter found in the given problem set that can sufficiently help to build an accurate predictive model [23].

3) Classification tasks

Binary classification is a classification with two possible outcomes. For example, weather forecast (it will rain or not), spam or fraud detection (predict whether an email is spam or not). Multi-label classification is a classification task with more than two possible outcomes. For example, classify academic performance of students as excellent or good or average or poor. In classification, a sample can even be mapped to more than one tag labels. For example, a sample news article can be labelled as a sport article, an article about some player, and an article about a certain venue at the same time.

4) Logistic linear regression (for classification problems)

These are statistical models by which a logistic curve is fitted to the dataset [21], [25]. This is applied to when the dependent variable or target variable is dichotomous. There is well defined probabilistic interpretation, and the model can be updated to take new data simply through the application of gradient descent method. As it returns probability, the classification thresholds can easily be adjusted. Fewer assumptions to no assumptions are applied on the distribution of the independent variables, no linear relationship between the predictors and the target variable must be assumed, it can handle interaction effect and power terms [26], [27]. A large sample size is required to maintain stable results.

It is widely used as a classification algorithm, when the dependent variable is in a binary format, so to predict the outcome of a categorical dependent variable. The outcome therefore becomes discrete or categorical in nature [21]. Here discrete means values should be binary, or it can have just two outcomes: either 0 or 1, either true or false, either yes or no, or either high or low. In logistic regression we don't need the value below 0 and above 1. In logistic regression, we must predict categorical variables and solve classification problems.

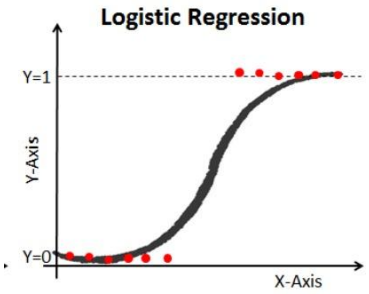

Fig. 1 Sigmoid curve **[28]**

In the above diagram there is a sigmoid curve which converts any value from negative infinity to infinity to binary values which a logistic regression model need. The logistics Regression Equation is derived from the straight-Line equation.

Advantages

- Implementation is quite simple.
- Mathematical proficiency
- Proficiency in regulation
- Efficient with regards to training
- No scaling needed for input variables.
- Works efficiently with large dataset.
- Easily extend multiple classes
- Can be overfit with multi scale which is controlled by the technique known as regularization.
- Outputs are more calibrated than most models.

Disadvantages

- Cannot solve non-linear problems.
- Complex elation is difficult to handle.
- Good predictions require huge dataset.
- Duplicity of data can lead to wrong training parameters.

III. Methodology

A. Research design

This research work made use of primary and secondary data. The primary data was collected from laboratory labs where the hydrometallurgical treatment of cobalt was conducted. This investigation involved varying leaching parameters to generate a yield of cobalt. The leaching parameters involved pH, temperature, acidity, PSD, Duration, Pulp D, Agitation, and Reducing agent. In the context of the secondary data, journal articles, conference papers and posters were used to generate the dataset used for this project. The combination of primary data and secondary data created a huge dataset that has been used to evaluate the use of machine learning in the leaching of cobalt.

Logistic linear regression was chosen for this project, based on its relevance to classification problems. It works well with linear data. For thus research, Jupyter notebook was used to write codes and perform logistic linear regression technique, the programming language used was Python. The codes were generated through consultation with various sources and machine learning domains, such as Scikitlearn, Pandas, Tensorflow and Kaggle. The interpretation of the generated results was based on domain knowledge (extraction of cobalt through leaching). Various steps were carried out for successful implementation of the model.

B. Data Processing

Data Frame

- 176 datapoints
- 177 rows
- 13 features
- 12 independent variables (pH, Temp, PSD, Acidity, Duration, Pulp D, Agitation, Reducing agent, Impurity ratios (Co: Cu, Cu: Fe, Co: Fe, Co: Mn)).
- 1 dependent variable (yield of cobalt)

Fig.3 shows a glimpse of how the dataset used for this research is structured as described above.

Run	pH	Temp	PSD	Acidity	Duration	Pulp D	Agitation	Reducing agent	Co:Cu	Cu:Fe	Co:Fe	Co:Mn	Yield of Co
1	8	25	103	70	120	30	800	5.08	0.15	1.51	0.23	3.20	82.76
2	8	60	106	70	120	26	600	2.5	0.15	1.51	0.23	3.20	91.23
3	8	47	75	70	120	20	600	2.5	0.15	1.51	0.23	3.20	82.87
4	4	60	75	70	120	20	800	5.08	0.15	1.51	0.23	3.20	86.72
5	4	60	84	70	120	30	600	3.79	0.15	1.51	0.23	3.20	77.38
6	4	25	75	70	120	20	600	5.08	0.15	1.51	0.23	3.20	78.6
7	6.3	48	75	70	120	20	600	2.5	0.15	1.51	0.23	3.20	87.44
8	8	60	106	70	120	20	800	5.08	0.15	1.51	0.23	3.20	91.51
9	5.4	50	87	70	120	25	732	2.5	0.15	1.51	0.23	3.20	47.43
10	4.4	25	106	70	120	20	800	2.5	0.15	1.51	0.23	3.20	80.5

Fig. 2 Dataset used for this study.

For a good model performance, some data cleaning techniques were implemented, such as feature engineering, feature scaling, normalization of the features, outlier identification, dealing with missing values. The code in the appendix shows how the above-described process was implemented. The first thing that was done was to convert the dependent variable, which is the yield of cobalt as seen in Fig. 4 into a binary system (true or false). The breaking point was the threshold, which describes the two classes for prediction. Two classes were generated, (High yield and low yield). With high yield representing the yield of cobalt above the threshold value (75%) and low yield represents the yield of cobalt below the threshold. The threshold value was decided because cobalt is mainly extracted as a by-product therefore having a yield of 75% and above is good.

```
0       True
1       True
2       True
3       True
4       True
       ...
171     True
172     True
173     True
174    False
175     True
Name: Yield of Co binary, Length: 176, dtype: bool
```

Fig. 3 Binary conversion of the target variable (Yield of Cobalt)

C. Feature Selection

Identifying most relevant features (independent variables) for the prediction. SelectKBest with f_classif metric was used for fearture selection. Also, the implementation of L1 regularization. This is seen on the appendix on the generated code. Random forest feature importance of selected features was analyzed using the feature important score. The best feature score is 1, if a feature has a score closer to one, it is important towards the prediction analysis. It is to be noted that based on the type of dataset and how it structured, this might contradict with domain knowledge. It is best to consider domain knowledge when it comes to feature selection. For this study, feature selection was generated to see how the model performs towards this task.

D. Model training

The model training was done on 80% of the dataset and the remaining 20% was used for testing purposes.

E. Model performance validation

After the generation of a model, its performance must be evaluated to see how it performs towards the specific dataset. The validation is carried out using different evaluation metrics. Often, the model performance is not valuated with the same dataset used for training the model, this is to avoid overfitting the model and incorrect predictions as it is bound to perform very well if the predictions are carried out on the same dataset used for training.

The first thing that is obtained is the confusion matrix, this describes the classification errors obtained by the model. The confusion matrix is a 2 × 2 matrix with numerical values True Positive (TP), True Negatives (TN), False Positives (FP), and True Negatives (TN), which are the result of the classified cases, where TP is the sum of the true positive cases, TN is the true negatives, FP represents the false positives, and FN corresponds to the false negatives (ref).

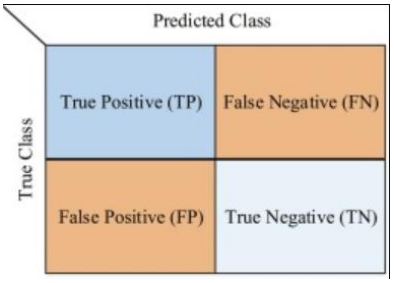

Fig. 4 Confusion matrix for a binary classification

The measures of merit are used to assess the quality of the predictive model developed and are calculated based on the data from the confusion matrix and results from the training of the classification model.

1) *Accuracy* -Accuracy (acc) corresponds to the ratio of correctly classified samples from all the samples in the dataset. This indicator can be calculated with the confusion matrix data, according to Equation (1)

$$ACC = \frac{TP+TN}{TP+TN+FN} \quad (1)$$

2) *Precision* -Precision (p) is the proportion of true positives (TP) 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 Equation (2),

$$P = \frac{TP}{TP+FP} \quad (2)$$

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 Equation (3),

$$r = \frac{TP+\ FN}{TP+FP} \quad (3)$$

4) ROC (Receiver operating characteristics) and AUC (Area under curve)

- is a performance measurement for classification problems at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.

IV. Results and Discussion

A. Feature Importance

In this section, the model assessed the relationship between the features selected by the model, based on the weight they carry towards the target variable. The algorithm implemented for the analysis was random forest importance feature. It analyzed based on the selected features as implied in the methodology. Figure depicts the results from the analysis.

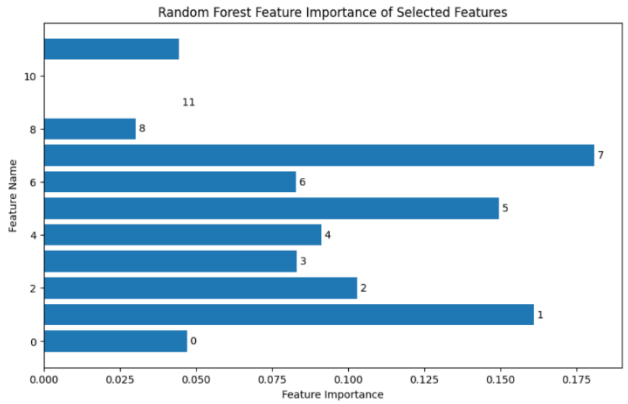

Fig. 5 feature importance levels; Feature Names: 0 = pH, 1 = Temperature, 2 = PSD3 = Acidity4 = Duration5 = Pulp D, 6 = Agitation, 7 = Reducing agent, 8 = Ratio (Co: Cu), 11 = Ratio (Co: Fe)

The feature important analysis showed that the most weighted feature having the most significant impact on the predicted output, cobalt yield with an importance score of 0.175, this is an indication that the reducing agent has a huge impact towards the yield of cobalt. As it has stipulated in the literature, the use of the reducing agent enhances the dissolution of cobalt, particularly from an oxide ore. This thus indicates the ability of classification algorithm to predict the significance each feature holds or has towards the output. As figure 3 stipulates, the selected features have an important role in the leaching of Cobalt.

B. Comparison between actual and predicted outcomes

The implementation of feature engineering plays a huge role in the performance of the logistic linear regression model. The 'Yield of Cobalt' was converted into a binary system (Yield of cobalt binary) to best suit the models need, so to enhance its performance. As indicated in the methodology, the classification task held was based on two categories, high yield and low yield of cobalt. The break point was imposed by the threshold, which is a distinctive parameter in the implementation of a logistic linear regression model. The threshold was set to 75 % (based on the idea that cobalt is recovered as a byproduct, thus a yield of 75% and above is good), this indicates that for a specific datapoint the yield of cobalt could be either above 75% indicating a class of 'High yield' or below 75% which indicates 'Low yield' depending on the leaching parameters complementing that specific datapoint. With 'True' indicating the yield of cobalt above and 'False' below 75%. Table 4 indicates a glimpse of what the predictions were like. The predictions are for the target variable, 'yield of cobalt binary'.

TABLE I
COMPARISON BETWEEN ACTUAL AND PREDICTED OUTCOMES

Datapoint	Actual	Predicted
19	True	True
45	True	True
139	True	True
30	True	True
67	False	False
16	True	True
119	True	False
172	True	True
109	True	True
140	False	False

The predictions were correct for most instances as indicated in figure 4. Datapoint 19, shows that before the model prediction the ‘yield of cobalt’ was above 75% (indicated by true), meaning the leaching parameters at that datapoint lead to a high yield outcome of cobalt, comparing the actual to the predicted for the same datapoint, it is visible that the model made a correct prediction. Affirming to what the actual indicates in terms of the yield of cobalt generated from the leaching parameters. But, as for datapoint 119, the situation is not the same. Observing datapoint 119, before the model prediction, the actual outcome was ‘True’, but the model predicted it to be ‘False’. This indicates that the models’ evaluations were off, this could imply that the datapoint appeared to have two instances, ‘High yield’ or ‘Low yield’.

C. Model Evaluation

The results were evaluated using performance metrics (Accuracy, precision, and recall), as indicated by Table II.

TABLE II
MODEL VALIDATION RESULTS

	Metric	After Model Training
0	Accuracy	0.750000
1	Low Yield (precision)	0.454545
2	Low Yield (recall)	0.625000
3	High Yield (precision)	0.880000
4	High Yield (recall)	0.785714

As depicted in table II the model performed well, as it performed an overall accuracy of 75%. The precision for the "Low Yield" class is approximately 0.4545, which indicates that about 45.45% of the predictions for "Low Yield" are correct. The recall for the "Low Yield" class is 0.625, meaning that the model correctly identifies 62.5% of the instances of "Low Yield" in the dataset. The precision for the "High Yield" class is 0.88, indicating that about 88% of the predictions for "High Yield" are correct. The recall for the "High Yield" class is approximately 0.7857, meaning that the model correctly identifies about 78.57% of the instances of "High Yield" in the dataset.

The predictions on the ‘High Yield’ instances are reasonably good. The concern is with the ‘Low Yield’ predictions, this is an indication of the dataset itself. Most of the datapoints were obtained through secondary sources, in these sources the emphasis was on the optimization of the cobalt recovery from leaching by manipulating various parameters, this therefore affects the data frame. The optimized (from secondary sources) instances indicate that most of the datapoints depict a high yield of cobalt, thus low yield of cobalt holds less weight on the dataset. Thus, the model is most likely to make correct predictions on high yield class of cobalt compared to low yield class of cobalt.

D. Confusion matrix

The imbalance in the dataset has been established in the previous section, whereby it seemed that there are more high yield instances in the dataset as compared to the low yield instances, henceforth it is most likely that the predictions lean towards high yield class. As it has been established that the model performs poorly when predicting instances of the ‘Low yield’ class. The confusion matrix visualizes the imbalance of the prediction of the two classes.

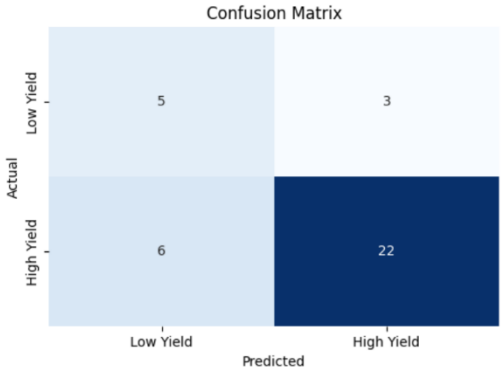

Fig. 6 Confusion Matrix results

Figure indicates that the number of predictions on ‘True positive’ is 5, this means that about five times the actual positive values are equal to the predicted positive, therefore positive values were predicted and are correct. As for the ‘True negative’, about 22 iterations of negative values are equal to the predicted negative values, thus the predicted negative value is negative and correct. Looking at the False positive, the model wrongly predicted negative values as positive 3 times, thus the predicted value was negative and its positive. The same applies to the false negative, the model predicted wrongly negative values as positive 6 times, thus the predicted value is negative, but the actual is positive. Although this might not be detailed, it shows that the model had some problems when it came to predicting the low yield class.

The ROC (Receiver operating characteristic) curve depicts the significance of the threshold towards the predictions. This is to evaluate how the chosen threshold is most favorable for this classification problem.

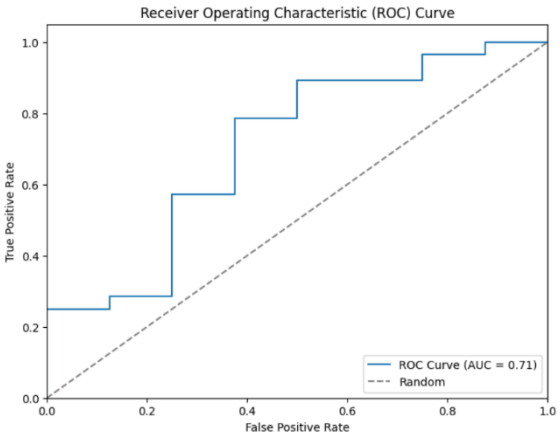

Fig. 7 ROC and AUC results

The model achieved an AUC (Area under curve) value of 0.71, which therefore details that the model’s predictions were good. The model successfully predicted high yield situations. The curve is leaning towards the true positive rate. With a significant amount of false positive.

E. Overall discussion

The chosen model performs well for the classification problem. The model performs well when it comes to predicting the ‘High yield’ instances as compared to the ‘Low yield’ instances. This is merely due to the size of the dataset and its shape. The significant amount of data representing the ‘Low yield’ instances dictates the performance of the model when predicting for that class, as less data is picked by the model to train itself for the ‘Low yield’ instances.

REFERENCES

[1] B. Mbuya., P. Ntakamusthi., M.B. Kime, L. Zeka., G. Nkulu., A. Mwamba., & A.F. Mulaba-Bafubiandi “Metallurgical Evaluation of the Leaching Behavior of Copper–Cobalt-bearing Ores by the Principal Component Analysis Approach: Case Study of the DRC Copperbelt Ore Deposits,” Journal of Sustainable Metallurgy, vol. 7, no. 3, pp. 985–994, Sep. 2021, doi: 10.1007/s40831-021-00389-5. https://doi.org/10.1007/s40831-021-00389-5

[2] B. Mbuya and A. F. Mulaba-Bafubiandi, “Predicting Optimized Dissolution of Selected African Copperbelt Copper-cobalt-bearing Ores by Means of Neural Network Prediction and Response Surface Methodology Modeling,” Process Integration and Optimization for Sustainability, 2023, doi: 10.1007/s41660-023-00312-3. https://doi.org/10.1007/s41660-023-00312-3

[3] B. I. Mbuya, M. B. Kime, and A. M. D. Tshimombo, “Comparative Study of Approaches based on the Taguchi and ANOVA for Optimising the Leaching of Copper–Cobalt Flotation Tailings,” Chem Eng Commun, vol. 204, no. 4, pp. 512–521, Apr. 2017, doi: 10.1080/00986445.2017.1278588. https://doi.org/10.1080/00986445.2017.1278588

[4] E. Cavallero, B. Murray, A. Professor, and G. Kim, “Forecasting the Effects of Battery Recycling on the Global Cobalt Market.”

[5] S. Y. Tshipeng, A. TshamalaKaniki, and M. B. Kime, “Effects of the Addition Points of Reducing Agents on the Extraction of Copper and Cobalt from Oxidized Copper–Cobalt Ores,” Journal of Sustainable Metallurgy, vol. 3, no. 4, pp. 823–828, Dec. 2017, doi: 10.1007/s40831-017-0149-x. https://doi.org/10.1007/s40831-017-0149-x

[6] M. H. M. Mwanat and K. B. Kasongo, “Cobalt Dissolution from Concentrate in Sulfuric Acid—Ferrous Sulfate System: Process Parameters Optimization by Response Surface Methodology (RSM),” Journal of Sustainable Metallurgy, vol. 7, no. 4, pp. 1838–1851, Dec. 2021, doi: 10.1007/s40831-021-00460-1. https://doi.org/10.1007/s40831-021-00460-1

[7] B. Mbuya, J. Meta-Mvita, and A. F. Mulaba-Bafubiandi, “Bayesian statistics study of a sustainable dissolution of cobalt-bearing minerals from Cu-Co ores,” Canadian Journal of Chemical Engineering, 2022, doi: 10.1002/cjce.24817. https://doi.org/10.1002/cjce.24817

[8] S. Biswas and A. Floribert Mulaba-Bafubiandi, “Extraction of Copper and Cobalt from Oxidized Ore using Organic Acids Desulphurization of South African coal using low power microwave energy View project Let’s Save the World and Humanity View project,” 2016. [Online]. Available: https://www.researchgate.net/publication/305851241

[9] F. K. Crundwell, N. B. du Preez, and B. D. H. Knights, “Production of cobalt from copper-cobalt ores on the African Copperbelt – An overview,” Miner Eng, vol. 156, Sep. 2020, doi: 10.1016/j.mineng.2020.106450. https://doi.org/10.1016/j.mineng.2020.106450

[10] S. S. Afolabi, M. O. Zakariyah, M. H. Abedi, and W. Shafik, “A survey on cobalt metallurgical processes and its application,” Journal of the Indian Chemical Society, vol. 98, no. 11. Elsevier B.V., Nov. 01, 2021. doi: 10.1016/j.jics.2021.100179. https://doi.org/10.1016/j.jics.2021.100179

[11] N. Peeters, K. Binnemans, and S. Riaño, “Recovery of cobalt from lithium-ion battery cathode material by combining solvoleaching and solvent extraction,” Green Chemistry, vol. 24, no. 7, pp. 2839–2852, Mar. 2022, doi: 10.1039/d1gc03776e. https://doi.org/10.1039/D1GC03776E

[12] S. Y. Tshipeng, A. TshamalaKaniki, and M. B. Kime, “Effects of the Addition Points of Reducing Agents on the Extraction of Copper and Cobalt from Oxidized Copper–Cobalt Ores,” Journal of Sustainable Metallurgy, vol. 3, no. 4, pp. 823–828, Dec. 2017, doi: 10.1007/s40831-017-0149-x. https://doi.org/10.1007/s40831-017-0149-x

[13] M. H. M. Mwanat and K. B. Kasongo, “Cobalt Dissolution from Concentrate in Sulfuric Acid—Ferrous Sulfate System: Process Parameters Optimization by Response Surface Methodology (RSM),” Journal of Sustainable Metallurgy, vol. 7, no. 4, pp. 1838–1851, Dec. 2021, doi: 10.1007/s40831-021-00460-1. https://doi.org/10.1007/s40831-021-00460-1

[14] S. Y. Tshipeng, A. TshamalaKaniki, and M. B. Kime, “Effects of the Addition Points of Reducing Agents on the Extraction of Copper and Cobalt from Oxidized Copper–Cobalt Ores,” Journal of Sustainable Metallurgy, vol. 3, no. 4, pp. 823–828, Dec. 2017, doi: 10.1007/s40831-017-0149-x.

[15] M.-B. Kime and A. Floribert Mulaba-Bafubiandi, “Value Recovery from Mukondo Mine Low-Grade Cobalt Ore by Heap Leaching and Solvent Extraction Gravity concentration of copper-cobalt ores View project CALL FOR SUBMISSIONS: Case Studies in Chemical and Environmental Engineering (ELSEVIER) Special Issue ‘Biotechnological Advances for Resource Recovery and Bioremediation: Sustainable Practices’ (CSCEE-2021) View project”, doi: 10.13140/RG.2.2.35313.28001.

[16] B. Mbuya et al., “Metallurgical Evaluation of the Leaching Behavior of Copper–Cobalt-bearing Ores by the Principal Component Analysis Approach: Case Study of the DRC Copperbelt Ore Deposits,” Journal of Sustainable Metallurgy, vol. 7, no. 3, pp. 985–994, Sep. 2021, doi: 10.1007/s40831-021-00389-5. https://doi.org/10.1007/s40831-021-00389-5

[17] B. Mahesh, “Machine Learning Algorithms-A Review Self Flowing Generator View project Machine Learning Algorithms-A Review View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review,” International Journal of Science and Research, 2018, doi: 10.21275/ART20203995.

[18] A. A. Soofi and A. Awan, “Classification Techniques in Machine Learning: Applications and Issues,” Journal of Basic & Applied Sciences, vol. 13, pp. 459–465, 2017. https://doi.org/10.6000/1927-5129.2017.13.76

[19] A. Singh, “Bharati Vidyapeeth’s Institute of Computer Applications and Management (BVICAM),” 2016.

[20] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, “Supervised Machine Learning Algorithms: Classification and Comparison,” International Journal of Computer Trends and Technology, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126. https://doi.org/10.14445/22312803/IJCTT-V48P126

[21] A. Choudhury, “The Role of Machine Learning Algorithms in Materials Science: A State of Art Review on Industry 4.0,” Archives of Computational Methods in Engineering, vol. 28, no. 5, pp. 3361–3381, Aug. 2021, doi: 10.1007/s11831-020-09503-4. https://doi.org/10.1007/s11831-020-09503-4

[22] A. Singh, “Bharati Vidyapeeth’s Institute of Computer Applications and Management (BVICAM),” 2016.

[23] P. C. Sen, M. Hajra, and M. Ghosh, “Supervised Classification Algorithms in Machine Learning: A Survey and Review,” in Advances in Intelligent Systems and Computing, Springer Verlag, 2020, pp. 99–111. doi: 10.1007/978-981-13-7403-6_11. https://doi.org/10.1007/978-981-13-7403-6_11

[24] J. E. T Akinsola, A. Jet, and H. J. O, “Supervised Machine Learning Algorithms: Classification and Comparison SQL Injection Attack (SQLIA) Detection and Prevention View project The Use Of BIG DATA in Mobile Analytics View project Supervised Machine Learning Algorithms: Classification and Comparison,” International Journal of Computer Trends and Technology, vol. 48, 2017, doi: 10.14445/22312803/IJCTT-V48P126. https://doi.org/10.14445/22312803/IJCTT-V48P126

[25] 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[26] J. T. McCoy and L. Auret, “Machine learning applications in minerals processing: A review,” Minerals Engineering, vol. 132. Elsevier Ltd, pp. 95–109, Mar. 01, 2019. doi: 10.1016/j.mineng.2018.12.004. https://doi.org/10.1016/j.mineng.2018.12.004

[27] B. Mahesh, “Machine Learning Algorithms-A Review Self Flowing Generator View project Machine Learning Algorithms-A Review View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review,” International Journal of Science and Research, 2018, doi: 10.21275/ART20203995.

[28]. Navalani,A. (2019, December 16). Understanding Logistic Regression in Python Tutorial. https://www.datacamp.com/tutorial/understanding - logistic-regression-python