National College of Ireland

National College of Ireland

Project Submission Sheet – 2020/2021

Student Name: …Shubham Garg, Ibrahim Rinub Babu, Iswarya Yogeashwaran……

Student ID: ……X19205295, X19207387, X20155034……

Programme: ……MSc. Data Analytics…… **Year:** …2021……

Module: ……Domain Application of Predictive Analytics……

Lecturer: ……Vikas Sahni……

Submission Due Date: ……August 20, 2021……

Project Title: ……PREDICTION OF SILICA IN IRON ORE USING MACHINE LEARNING TECHNIQUE……

Word Count: ……3312……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: …… Shubham Garg, Ibrahim Rinub Babu, Iswarya Yogeashwaran
……

Date: ……20-08-2021
……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

PREDICTION OF SILICA IN IRON ORE USING MACHINE LEARNING TECHNIQUE

ABSTRACT: Mining seems to be a sector in which expanding performance and effectiveness has been essential for revenue growth, because tiny changes in returns, pace, as well as performance could have a substantial influence. The overall purpose of this research would be to use data to forecast what percentage of adulteration is present with in ore concentrate. Since this pollutant has always been monitored once an hour, estimating that how so much silica would be in the ore concentration could really help benefit the technicians by providing pre-stage data to do measures. It is a necessary consequence, the technicians will be likely to access appropriate measure in earlier stage (minimize imperfection, once absolutely required) thereby contributing to sustainable development for surroundings. The machine learning models helps as predictive analytics technique to forecast the quality of mining by predicting the percentage of silica present in the ore. In this research, Random forest Regressor is being implemented to experiment the prediction and the results are assessed by the metrics such as precision, recall, Root mean square, and other metrics with an R square value of 0.90(90%).

Keywords: *Iron ore, Silica, quality management system*

1 INTRODUCTION

Iron ore made likely its most significant contribution during the Industrial Revolution in the mid of 1800s, when iron was initially utilized for the mass-produce of steel. More than forty percent of steel in that decade was used for developing infrastructure and building rail roots. Half of the British exportation was steel. According to The Economist, iron ore is the second most significant commodity after oil, and the creation of technologies to turn raw earth into steel is one of humanity's most brilliant achievements. When, it's come to resources the iron ore plays a very useful for every industry in the world. Although the iron ore can't be used directly by the people for their finalized product. As we know, we get raw material from the earth and the iron ore is one of raw material which we mine from the earth. The iron ore must go under the process where it will be treated with many other components. It processed with several chemicals along with this some physical substance. Under this series of process, we pull out many other minerals, but the percentage of iron is more as compared to other substances. Along with these substances we get some impurities as a waste know as Iron Ore Tailings (IOTs). Silica is one of the waste substances which extracted more and utilized less though make the increase of cost in the final product from the ore. According to the recent studies in the ore clearly shows that iron ore is used as a raw material in manufacturing steel as compared to others. Another study indicates that each tone has been produced from their twice to trice tones and in total around 130 tones has been produced annually, which plays a very important role for the GDP of a country apart from it around 1.52 million is wasted per year. Considering this fact, the people make some laboratory test before starting the mining will attain the determined standard for the goods and machine learning is playing very important role for it through this, we can predict the percentage of waste in the iron ore, and it will save the initial and laboratory cost. In early research they try to predict on single dependent variable with two or at most three-independent variable through Multi Linear Regression but here we will be taking 19 Independent variables with single dependent variable that is silica and applying Random Forest Regression.

A. OBJECTIVE:

To predict the % of impurity That would help to engineers to act in predictive and optimized way to improve the iron ore quality and to escalate the return of investment (ROI).

- Examine various methods for forecasting the presence of impurities in ore concentrates.
- Building a model that can precisely predict the amount of silica is present in the ore concentrate.

B. RESEARCH QUESTION:

How effectively can the % of impurity of each ore be predicted to improve the Return of Investment (ROI) in mining industry?

C. HYPOTHESIS:

The impact of poor quality of the iron ore will not result in the reduction of return of investment (ROI)

2. RELATED WORKS

In the research [1] the author proposed the multi target regression model in which he state that there can be more than one dependent variable. It helps to understand the mapping to output dependent variable from input variable together. The research also shows that Multiple target regression is carried out for the forecasting quality during the mining and identifying the amount of silica in the mining process of iron ore. More than one algorithm has been used and compared to get the best model out of it the models which were performed are AdaBoost,

decision tree, k-Nearest neighbor. This experiment is good because of estimating the two values simultaneously.

This research paper [7] talks about the environmental system, as in the developing era of mining, environment has become an issue for all over the world as increasing the global warming and the pollution is increasing day by day, so they use the correlation analysis between the sources, environment and social economic. They develop the system which will be considering all three factors and provide a single mine environmental quality. It can be assigned by standardizing the values of all three and then the correlation analysis processed. And when experimenting this algorithm, they find a reliable mine environmental evaluation system that is Antu county, Jilin Province Mining Ltd, etc. as an example.

In this paper [3], they study and experiment about the real time estimation of impurities after the mining before the final product. The process in between is flotation. Prediction during the process of flotation give the engineers for getting the major effect on quality of the final product, as if they know the portion of impurity in the ore they can directly work accordingly and run the flotation more efficiently. For this process of finding the impurity in the ore they perform some machine learning algorithms to find the best out of them which are Long short-term memory, Gated recurrent unit. They compared on several factors like, mean square error, error percentage, mean absolute error, root mean square. On comparing they got the good result from LSTM with below 9% error.

This research [4] focus on the Philippines where they don't have many resources because of which they don't have the good equipment with whom they can measure the quality of surroundings like air, water, and soil elements. These are the basic element through which we can ensure the safety for miners. To solve these obstacles the mobile electronic sensors are being used, to measure the water from nearby mining sites, air. They track them through Atomic absorption spectroscopy analysis. The data recorded in every two months. After that we take our usual steps of modeling and then they apply conditional inference tree. Through this model we obtain a column of decision where we get the classified result which will be good, bad, or unknown.

In this publication [2], Distinctive regression models which use Random Forest, AdaBoost, k-Nearest Neighbors, and Decision Tree algorithms totally separate inside the context could be evaluated by comparing in the observational research to select the optimum model. The evaluation metrics has coefficient of determination (R square). Some research predicts iron concentrate and silica concentrate individually. Nevertheless, these two values do have strong correlation, one such paper makes a novel contribution to the world by measuring them together.

To eliminate the occurrence in this research [5], researchers investigated the use of artificial neural networks. This enabled the detection accuracy to fulfil the criteria of innovation while avoiding the correlation coefficients effect. It resulted in an increase in the consistency of logistical system designs.

3 METHODOLOGY

The strategy for predictive analytics is carried out by the below shown Figure-1.

A. DATA SELECTION:

The Quality mining Prediction Dataset is collected from a mining industry, which is available as a CSV file form in Kaggle website. This is available for public access without any restrictions. This data is the process held in the period from March to September in the year of 2017. The dataset comprises of numerical variables, which is about 24 columns in total. The total rows of the data are about 736282. The first column is date and time of the process, in which some of the m are counts of each second and some are hourly basis. The second column is the iron ore quality measurement before it goes into the floatation process. The fourth to eight columns are the significant columns which influences the iron ore quality. The nineth column to twenty second column are the significant columns which are the process information (like air flow and levels) involved in the floatation. The quality of ore at the last- stage is measured from lab, which are the last two columns. The independent variable is the last column, that is being predicted. It is the percentage of silica (filth) present in the ore concentration.

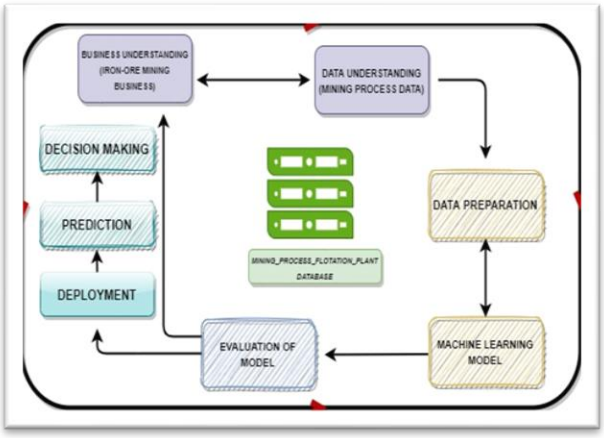

Figure-1

B. DATA PRE-PROCESSING

1. MISSING VALUES:

The missing values does not help to obtain the prediction at good level. The missing values in the attributes can be filled in two ways- one is by using mean or mode values and the other one is deleting the column which has missing values. In this dataset, there is no missing values present. The date colu mn has ignored, because it has no dependence on the dependent Variable.

2. MULTICOLLINEARITY:

The correlation is the statistical mutual relationship between two variables. It will be between +1 to -1. The multicollinearity is the problem that if two independent variables have correlation with one another. In this dataset, the columns such as Flotation Column 01 Air Flow and Flow Flotation Column 03 Air Flow, Flotation Column 05 Level and Flotation Column 07 Level, and Flotation Column 02 Level and Flotation Column 01 Air Flow have correlation between each other. Therefore, columns like Flotation Column 01 Air Flow, and Flotation Column 05 Level are removed to avoid the multicollinearity

issue. The correlation between all the variables is represented in the Figure-2.

3. FEATURE IMPORTANCE:

The major concern involved in selecting the attributes, because the unnecessary attributes will bias the results to predict. The feature importance is the method to select the attributes which is indeed and appropriate to predict the dependent variable. It is calculated by getting scores for all the columns and the columns with high score will be selected to proceed with building the models. Because those columns will have higher effect on the building model. In this dataset, the 20 columns except the columns include date, Flotation Column 01 Air Flow, Flotation Column 05 Level and Floatation column 07 Level are selected to build model.

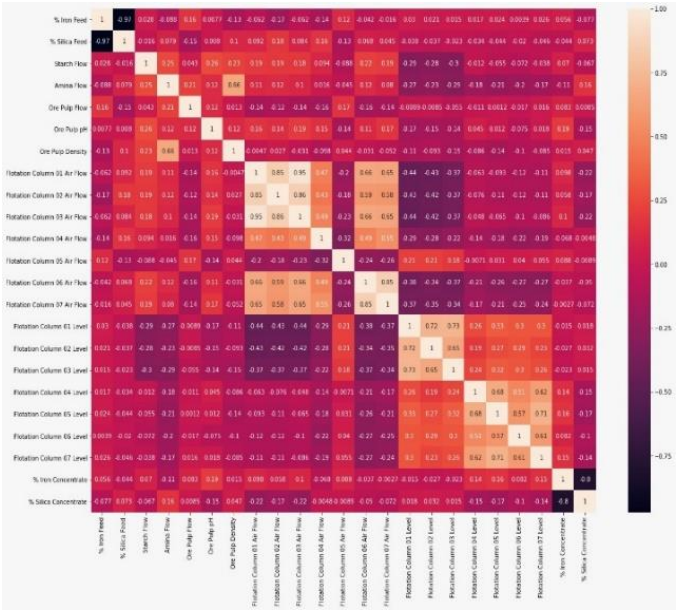

Figure-2

C. TECHNIQUE APPLIED

The machine learning has two types of problems such as classification and regression. The classification is the algorithm in which the dependent variable will have two or more categories, whereas regression is the problem where the dependent variable will be numerical values. This is the study used to predict the quality of ore by predicting the amount of silicae percentage. In this research, Random Forest Regressor is applied to predict the silica percentage. Random Forest Regressor is one of the famous methodologies for forecasting continuous variable, because it is simple and gives higher accuracy. This algorithm is a supervised learning which learns to maps the inputs and outputs during training. The entire dataset is splitted into train and test data. The proportion of train and test data is 0.7 and 0.3. This is splitted like, 70 percentage of the data is train data and remaining 30 percentage is testing data.

MODEL SELECTION JUSTIFICATION:

Business forecasting frequently employs random forest algorithms to produce machine learning predictions and decision making. The random forest employs numerous decision trees to conduct a more comprehensive examination for the aggregated outcome of a single data collection process. Because of its simplicity and excellent accuracy, it is among the most used algorithms for regression issues such as forecasting continuous outcomes. In this study, random forest is used to predict iron ore impurity, which helps reduce the number of ore that falls to tailings by reducing silica in the ore concentrate. As a result, the quality of the iron ore improves, which helps to determine the firm's success in a variety of ways.

4. MODEL EVALUATIONS AND RESULTS

A. EVALUATIONS

Evaluation metrics are the best way to evaluate the models. The regression problems are assessed by the evaluation metrics such as R square value, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Square Error (MSE), which are shown in the Figure-3. The values are visually represented in the Figure-4.

METRICS	VALUES
R Square Value	0.909013043
Mean Absolute Error (MAE)	0.009135923
Mean Square Error(MSE)	0.00124843
Root Mean Square Error(RMSE)	0.035333131

Figure-3

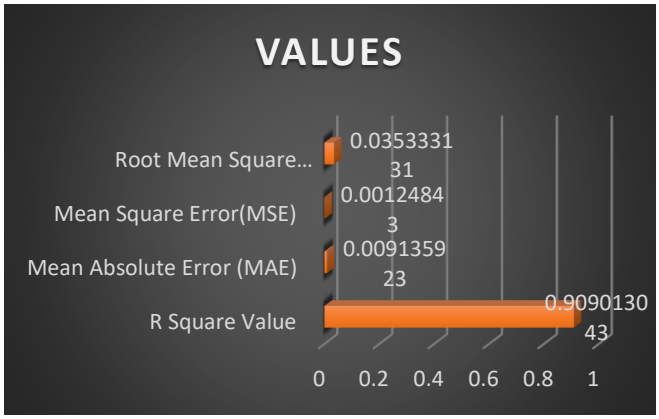

Figure-4

B. QUANTITATIVE RESULTS

The quantitative results interpreted in this research are,

- The Silica concentration has minimum of 1% while the Iron ore concentration has minimum of 63%.
- The Silica concentration is higher in September and April month (206.32% and 201.82%), while march and June month has lower percentage of 13.86% and 26.67% respectively.

The percentage of silica concentration predicted by using random forest, it is predicted value and then the actual value is compared in the graph to check the prediction accuracy in the Figure-5.

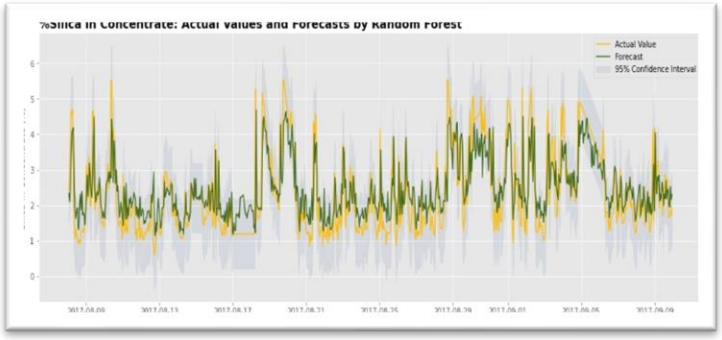

Figure-5

C. QUALITATIVE INTERPRETATION

Depends on the quality evaluation, the prediction and exploratory analysis made throughout using the datasets, for predictive analysis. The qualitative outcomes are listed below:

- The Sum of the Iron percentage and Silica percentage in the past data of the mining industry are 91.67% and 8.33% respectively.
- The production of Iron ore concentration is higher in the Second Quarter of the annual year, whereas it is lower in the First Quarter of the year 2017. The Third Quarter of the year is lesser than Second Quarter and higher than first quarter.

D. DATA DASHBOARDS AND VISUALIZATIONS

The Data Visualizations and Dashboards are the stage which showcase the insights of the business ideas. The Dashboards are the multiple combination of visualizations. The Figure-5 shows the percentage of Iron feed in the Floatation. It is splitted into high, middle, and low values to find the average percentage of Iron feed. There can be rise seen in the middle values of Iron feed.

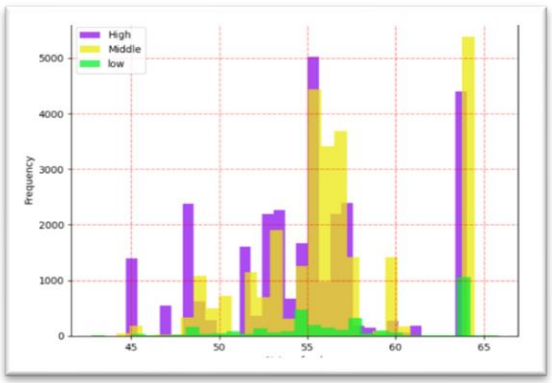

Figure-6

The Figure-7 shows distribution of the Silica concentration in the percentage. The distribution has highest impact in the range of 3.5%-4%.

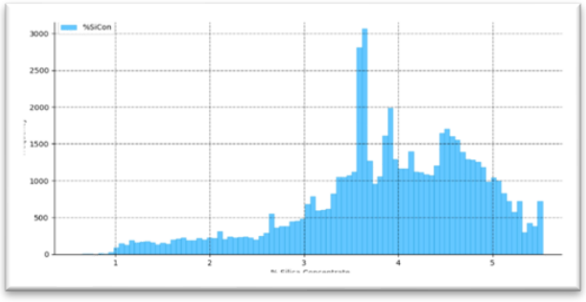

Figure-7

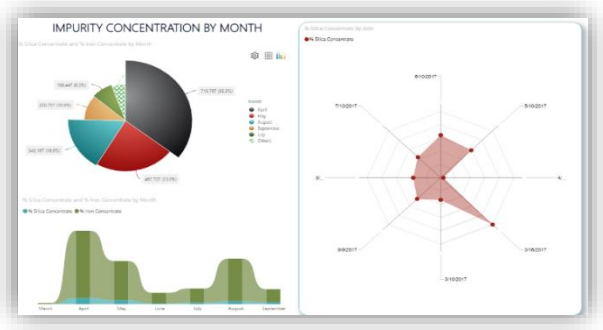

Figure-8

The Impurity concentration over Iron by the date and month is shown in the Figure-8, which is a dashboard. The April month has the bigger part of impurity concentration over Iron. The Figure-9 is a dashboard of multiple visualization about Floatation levels and airflow. The first one is the average value of floatation levels; floatation column level 3 is the highest one among all.

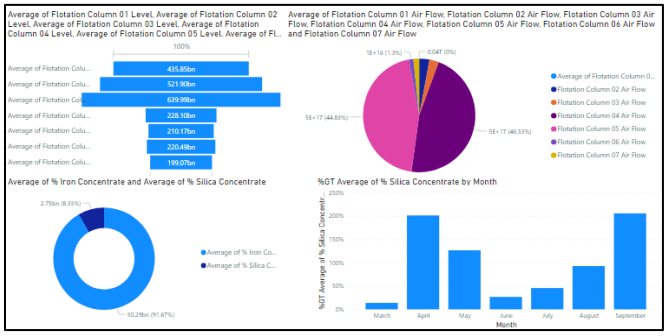

Figure-9

E. BUSINESS INSIGHT IDEAS

The quality of the product pertains to a product's ability to match user expectations. Most enterprises will fail if they cannot gain the trust of their customers. By integration of iron-ore mining datasets with the proposed machine learning leads to construct and customize the process of mining to achieve the production of high-quality iron ore. The business values are discussed below:

superior quality management system: From the insights gained by analyzing the historical iron-ore mining data and the outcome of predictive analysis performed by the proposed random forest model, the quality management system of the

mining industry under the regulation of ISO 9001:2008 standard can be planned and execution of deliver of the product in good quality can be done. Higher quality iron ore is utilized to develop the automobile, housing, and electronic industry, and other specialized infrastructure markets. The main objective for the high demand from the growing economies of countries such as China, Japan, and Europe, and the Middle East countries and India. The Company's entire shipments to China accounted for 61% of total exports. 19% of total exports were made around the rest of Asia and the remaining 12% were accounting for Europe.

Customer complaints and returns: Marketing researchers have consistently examined trade that gives high standard exports sustain higher replication of trade. This model will help to continuously improve and maintain the iron-ore quality. It also gives insights into when the quality is going to drop. This will be immense strength for planning the process flow.

Customers' trust: Numerous potential sales are missed when brands fail to engage with potential purchasers on a profound level. In contrast, if an enterprise gains consumers' trust and loyalty, they have more leeway to make judgments such as raising prices. But some of the actions that breaks vendors customer trust are, some corporations are purposely limiting supplies or boosting prices, EU regulators can punish them up to 10% of their annual global sales. one of the most crucial factors in gaining the trust of the target audience is to provide them with a high-quality product.

Word-of-mouth advertising: Launching campaigns to get folks talking about a product is a terrific approach to generate word-of-mouth recommendations. The response to complaints or compliments through the internet demonstrates that the business provides excellent customer service, which is another part of great product quality.

5 CONCLUSIONS

Predictive analytics is indeed an excellent technique to offer more clarity and insight into business choices. The design of the project carried out and then the methodology is implemented, predictive analytics is being applied to domain of the mining industry quality forecast. The Random Forest Regressor chosen in this research was assessed and used to predict the percentage of silica in the iron ore concentration. The methodology is applied in this dataset to predict the quality of iron. The model is built and evaluated by showing 90% of R square value. The model represents that it fits correctly. From the above insights and some past demonstration of Marketing researchers, businesses that provide high-quality goods receive greater repeat business. This will result in the improvement in ROI of the business.

REFERENCES

[1] A. A. CERAN and Ö. Ö. TANRIÖVER, "An experimental study for software quality prediction with machine learning methods," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020, pp. 1-4, doi: 10.1109/HORA49412.2020.9152918.

[2] A.Dogan, D. Birant and A. Kut, "Multi-Target Regression for Quality Prediction in a Mining Process," 2019 4th International Conference on Computer Science and Engineering (UBMK), 2019, pp. 639-644, doi: 10.1109/UBMK.2019.8907120.

[3] Haiyuan Qiu, "Study on the quality early-warning system of ecological environment in ore concentrated area," World Automation Congress 2012, 2012, pp. 1-4.

[4] J. Feng and Y. Qiao, "The quality prediction of iron ore pellets in grate-kiln-cooler system using artificial neural network," 2010 Sixth International Conference on Natural Computation, 2010, pp. 1906-1909, doi: 10.1109/ICNC.2010.5584645.

[5] M. Montanares, S. Guajardo, I. Aguilera and N. Risso, "Assessing Machine learning-based approaches for Silica concentration estimation in Iron Froth flotation," 2021 IEEE International Conference on Automation/XXIV Congress of the Chilean Association of Automatic Control (ICA-ACCA), 2021, pp. 1-6, doi: 10.1109/ICAACCA51523.2021.9465297.

[6] M. R. J. E. Estuar et al., "Towards building a predictive model for remote river quality monitoring for mining sites," TENCON 2015 - 2015 IEEE Region 10 Conference, 2015, pp. 1-5, doi: 10.1109/TENCON.2015.7373128.

[7] Y. I. Eremenko, D. A. Poleshchenko and Y. A. Tsygankov, "Prediction of Quality Indicators of Iron Ore Processing Operations Using Deep Neural Networks," 2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), 2020,pp.425-429,doi: 10.1109/SUMMA50634.2020.9280676.

[8] Y. Eremenko, D. Poleshchenko and Y. Tsygankov, "Neural Network Based Identification of Ore Processing Units to Develop Model Predictive Control System," 2019 XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP), 2019, pp. 121-124, doi: 10.1109/CSCMP45713.2019.8976790.

[9] Xuedong Wang, Guangjie Li and Bing You, "Evaluating the quality of mine environment based on rank correlation analysis," 2011 International Conference on Remote Sensing, Environment and Transportation Engineering, 2011, pp. 4952-4955, doi: 10.1109/RSETE.2011.5965423.