Industrial Internship Report on

"Quality prediction in mining process"

Prepared by

Hari Prasad N

Executive Summary
This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT). This internship was focused on a Quality Prediction in mining process project provided by UCT. We had to finish the project including the report in 6 weeks' time. My project was to explore real industrial data and help manufacturing plants to be more efficient The main goal is to use this data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions. Hence, they will be able to take corrective actionsin advance (reduce impurity, if it is the case) and also help the environment (reducing the number of ore that goes to tailings as you reduce silica in the ore concentrate). This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1 Preface

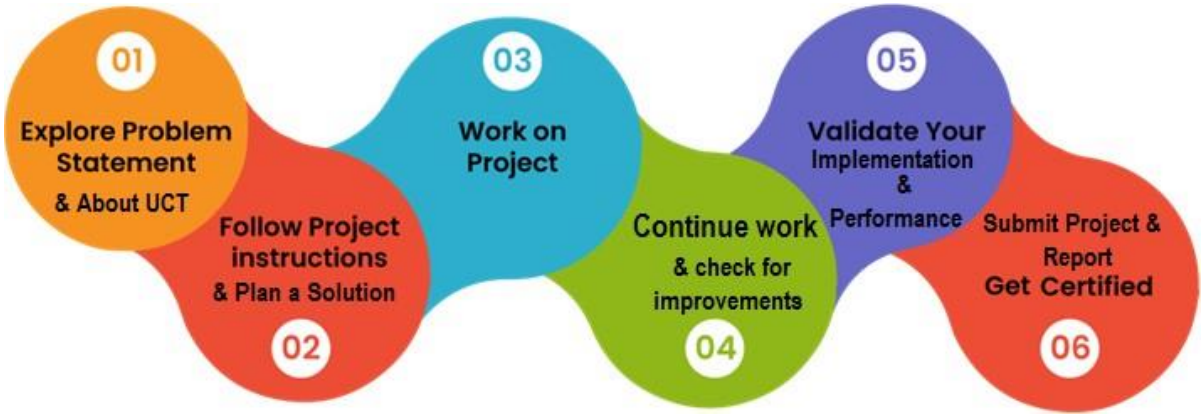
In the 6-week Internship program we were able to understand the concepts of DataScience and Machine Learning.

- Data Science: Introduction to the Data Science and we were able to gain the knowledge about the big data which is also created by the universe. The applications of data science in different domain areas and the usage of modern technology in our daily life.
- Machine learning: The origin of Machine learning and its development withthe definition and its application were explained. Relation of the Artificial Intelligence, Machine Learning and Deep learning concepts knowledge. Explanation of the machine learning working with the block diagram helpedus to explore our project.

1.1.1 Abstract

In the iron ore mining fraternity, in order to achieve the desired quality in the froth flotation processing plant, stakeholders rely on conventional laboratory test technique which usually takes more than two hours to ascertain the two variables of interest. Such a substantial dead time makes it difficult to put the inherent stochastic nature of the plant system in steady-state.Thus, the present study aims to evaluate the feasibility of using machine learning algorithms to predict the percentage of silica concentrate (SiO2) in the froth flotation processing plant in real-time. The predictive model has been constructed using iron ore mining froth flotation system dataset obtain from Kaggle. Different feature selection methods including Random Forest and backward elimination technique were applied to the dataset to extract significant features. The selected features were then used in Multiple Linear Regression, Random Forest and Artificial Neural Network models and the prediction accuracy of all the models have been evaluated and compared with each other. The results show that Artificial Neural Network has the ability to generalize better and predictions were off by 0.38% mean square error (mse) on average, which is significant considering that the SiO2 range from 0.77%- 5.53% -(mse 1.1%) . These results have been obtained within real-time processing of 12s in the worst case scenario on an Inter i7 hardware. The experimental results also suggest that reagents variables have the most significant influence in SiO2 prediction and less important variable is the Flotation Column.02.air.Flow. The experiments results have also indicated a promising prospect for both the Multiple Linear Regression and Random Forest models inthe field of SiO2 prediction in iron ore mining froth flotation system in general. Meanwhile, this study provides management, metallurgists and operators with a better choice for SiO2 prediction in real-time per the accuracy demand as opposed to the long dead time laboratory test analysis causing incessantloss of iron ore discharged to tailings.

How Program was planned



Learning Highlights:

i. Gained detail information about Data science and machine learning.

ii. Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.

iii. Concepts of model training and model Evaluation in machine learning as trained model, test model and validation model.

iv. AI technologies more demanding significance.

v. A good understanding of big data platforms like Hadoop, Statistical analysis is improved for this path.

I want to Thank Mr. Nithin Tyagi and Mr. Ankit for being a wonderful support network, your guidance and shared experiences have been invaluable. I appreciate everything you both have intimated the instructions and guidelines to complete my internship successfully.

I thoroughly enjoyed my internship this summer and now have very valuable experience under my belt. I know this will help when looking for jobs and needing references. Practical experience is the best and internships give students that hands on experience they need. I feel that quality internships are essential to develop key skills that we can't get in a classroom. Upskill campus provided us the practical experience and knowledge about the Data Science and Machine learning, helped to select the projects and completion of it. Thank you Upskill campus for providing me this internship opportunity.

2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

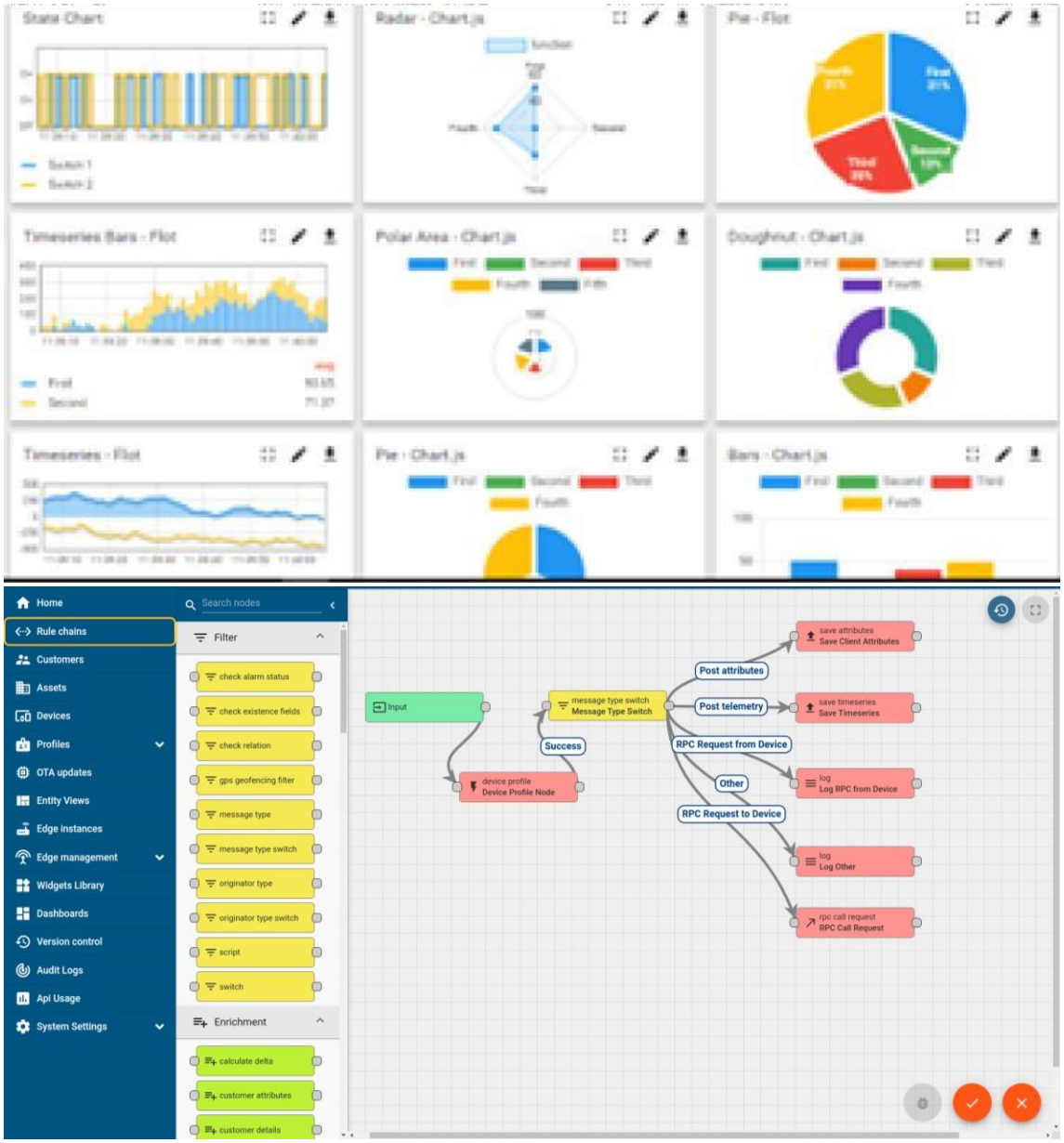
i. UCT IoT Platform (UCT Insight)

UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application(Power BI, SAP, ERP)
- Rule Engine



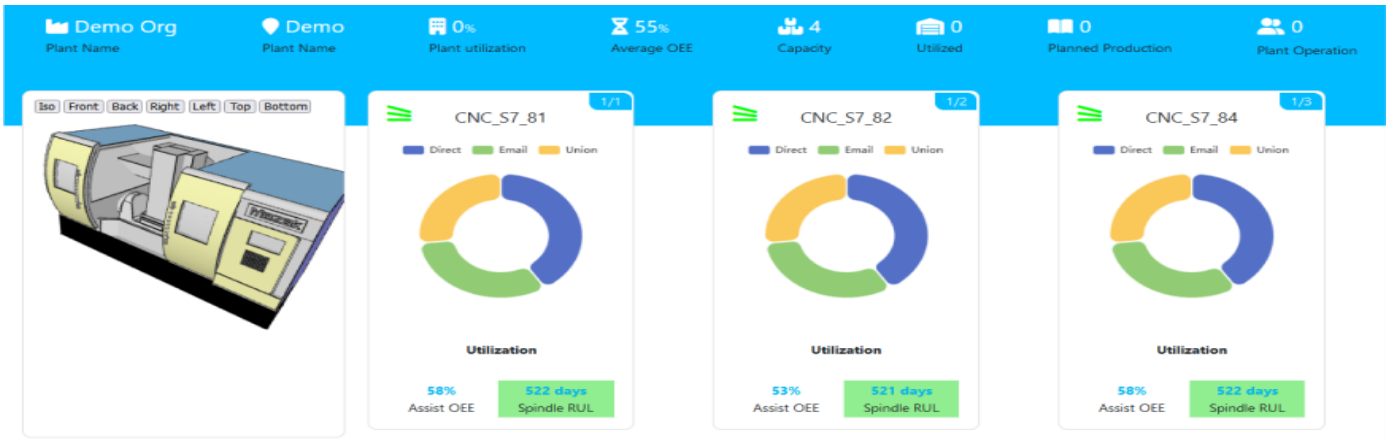
ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.



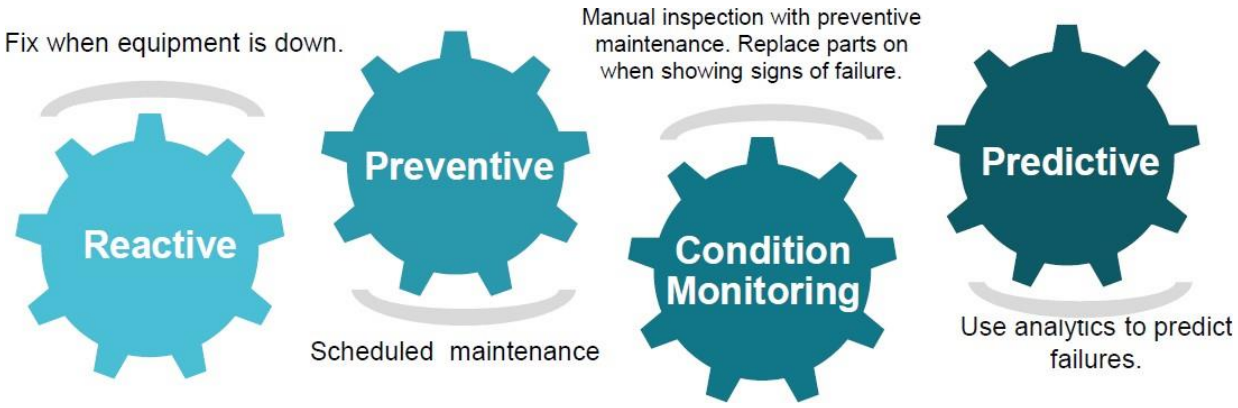
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

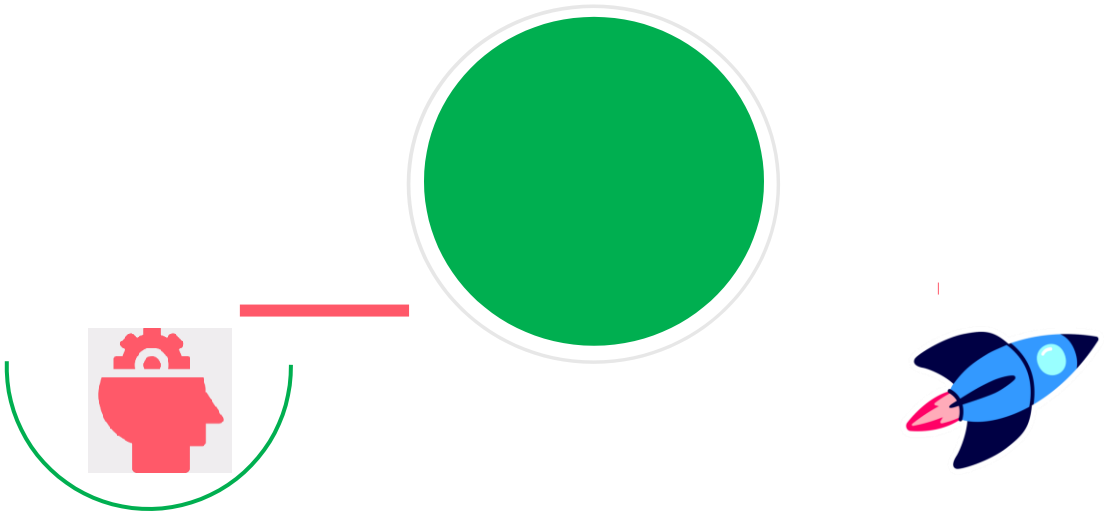
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

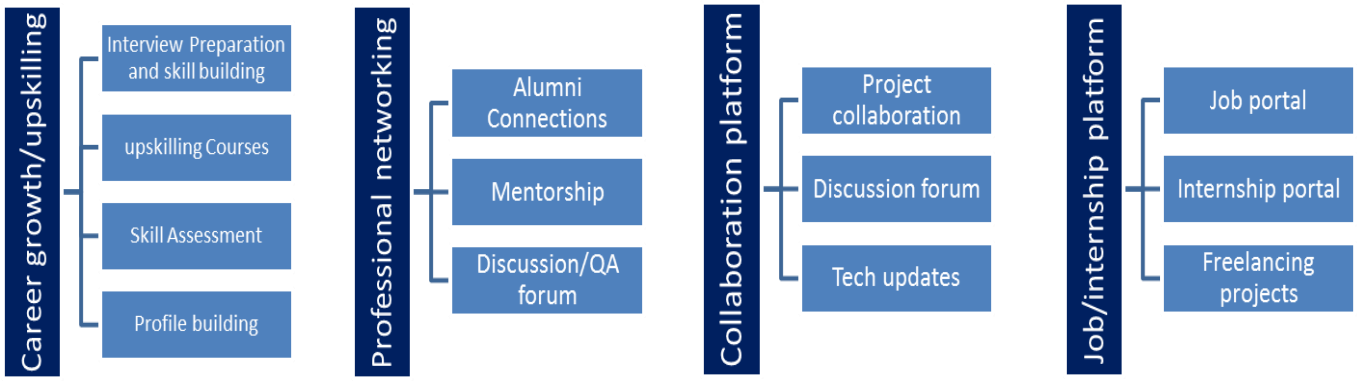
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/


Career growth/upskilling
Interview Preparation and skill building
upskilling Courses
Skill Assessment
Profile building
Professional networking
Alumni Connections
Mentorship
Discussion/QA forum
Collaboration platform
Project collaboration
Discussion forum
Tech updates
Job/internship platform
Job portal
Internship portal
Freelancing projects

2.3 Objectives of this Internship program

The objective for this internship program was to

- ☛ get practical experience of working in the industry.
- ☛ to solve real world problems.
- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.4 Reference

[1] Hume, N. (2014). End of the Iron Age. Financial Time. London, UK.

[2] Li, C., Sun, H., Bai, J., & Li, L. (2010). Innovative methodology for comprehensive utilization of iron ore tailings. Journal of Hazardous Materials, 174(1–3), 71–77.

[3] Dawson, P., & Koorts, R. (2014). Flotation Control Incorporating Fuzzy Logic and Image Analysis: IFAC Proceedings Volumes, 47(3), 352–357.

[4] Nakhaei, F., & Irannajad, M. (2015). Application and comparison of RNN, RBFNN and MNLR approaches on prediction of flotation column performance. International Journal of Mining Science and Technology, 25(6), 983–990.

[5] Bergh, L. G., & Yianatos, J. B. (2011). The long way toward multivariate predictivecontrol of flotation processes. Journal of Process Control, 21(2), 226–234.

[6] Jovanović, I., Miljanović, I., & Jovanović, T. (2015). Soft computing-based modeling of flotation processes – A review. Minerals Engineering, 84, 34–63.

[7] Parada, B. (2012). Study of liquid drainage in flotation foam. London, UK.

[8] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5).

3 Problem Statement

The main idea of the project is to predict the final quality of the iron concentrate. The input is an ore containing iron and silica, iron and silica feed, the output is silica and iron concentrate, the higher the value of iron concentrate, the better.

3.1 AIM: Explore real industrial data and help manufacturing plants to be more efficient

The main goal is to use this data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions. Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and also help the environment (reducing the amount of ore that goes to tailings as you reduce silica in the ore concentrate).

6.2 Objective:

I. To evaluate the feasibility of using machine learning algorithms to predict in real-time the percentage of silica concentrate of froth flotation processing plant.

II. Model selection: The project finds out which variable associated with iron ore extraction is statistically significant.

III. Estimate: The project will propose a model to predict percentage of silica concentrate.

4 Existing and Proposed solution

Over the past two decades, there has been an upsurge of academic research work within froth flotation process fraternity. Though, a significant number of the plant processing problems are being successfully modelled using machine learning algorithms but other unresolved issues and impediment still remain.

Notably amongst papers which have presented similar work on using Artificial Intelligence in the froth flotation process plant worth mentioning is the work of Dawson [3] in which the researcher explored the use of Fuzzy Logic to control the flotation stochastic process, resulting in increased grade and recoveries. The author further investigated the use of image analyser coupled with Fuzzy logic and reagents control to provide a starting point for expert knowledge to be utilized in order to monitor, evaluate and control grade and recovery. He however averred that, despite the success, there are still opportunities for further enhancements within the control jurisdiction of employing image analysing in the froth concentrate. The research concluded that availability of measurable features would be the basis for machine vision model to improve the froth flotation system problem.

Another author in the paradigm of froth flotation system has explored and discovered the most significant parameters that influence the flotation performance of lead mineral. Seemingly, the new model suggested that grinding time, flotation pH, for comparable collector, solid-in-pulp concentration and the increase of solid-in-pulp concentration have the most significant effect on the ore recovery and selective separation of lead mineral. He concluded that solid-in-pulp concentration was the most important parameter that influences the flotation of lead mineral [20].

In a recent study, advanced imaging systems based on Convolutional Neural Networks were employed to extract features from forth flotation plant, a case study of platinum flotation images at four distinct- grades. The extracted features were trained and compare with traditional texture feature extraction method. It was found that the results were competitive and nearly comparable [21]. Another study examined the used of several Neural Networks architecture models to predict metallurgical performance of the flotation column at Sarcheshmeh copper complex pilot plant. Apparently, 8 parameters were used namely: The chemical reagents dosage, froth height, air, wash water flow rates, gas holdup, Cu grades in the rougher feed, column feed and final concentrate streams concentrate streams were used for the simulation. The authors proposed Artificial Neural Networks (ANN) and Multivariate NonlinearRegression (MNLR) as the most robust model for predicting copper ore recovery and grade [4].

Furthermore, another study proposed artificial neural networks as most robust predictive model to estimate the percentage of silica concentration in iron ore mining plant with the aid of virtual sensor. In his work, 700K observations and 120 parameters including desliming variables which correspond to 10 seconds sampling of the plant process and laboratory variables [22].

The above review highlights the myriad of issues informing and contextualizing the novel approach in the froth flotation system. From this overview, it is apparent that predicting realtime percentage of silica

concentrate would be an essential component of the roll-out process. The chief objective, therefore in this thesis is to focus on using machine learning to predict online percentage of silica concentrate in the froth flotation process plant. More importantly, how to incorporate lagged values of silica concentration into the model.

4.1 Code submission (GitHub link)

https://github.com/HariPrasad2037/upskillcampus-/blob/main/silica%20prediction-checkpoint.ipynb

Report submission (GitHub link)
https://github.com/HariPrasad2037/upskillcampus-/blob/main/silica%20prediction_Hari_Prasad_N_USC_UCT.pdf

5 Proposed Design/ Model

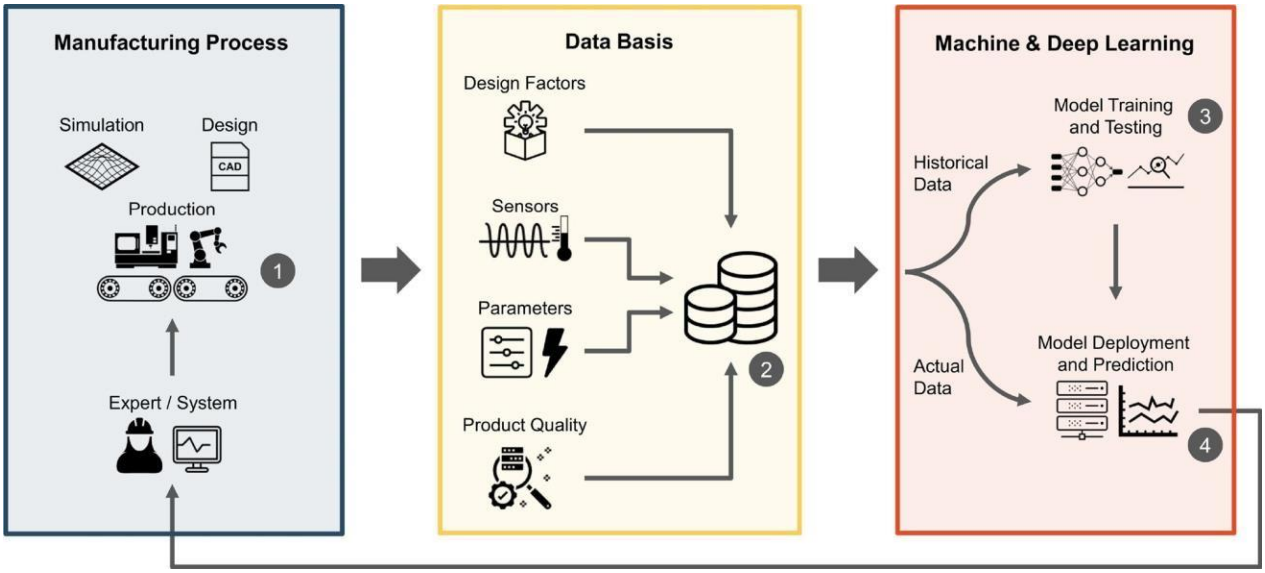


Figure 1 Predictive quality approach: for a selected manufacturing process (1), relevant process and quality data is collected (2) and used as a basisfor training a ML model (3). The trained model is used to perform quality estimations for decision support.

Our concern is whether

1. % Iron Concentrate is correlated with % Silica Concentrate

2.predict the % silica concentrate without using % iron concentrate.

3. If it is correlated and we can predict both % Iron and Silica concentrate at same time using power of ML and DL.

The dataset contains real values hence implementation of the regression model helps to answer the above concerns. To achieve it we follow these five steps.

1. Exploratory Data Analysis (EDA)

2. Pre-processing

3. Data splitting

4. Modeling

5. Evaluation

5.1 Source of Data

Kaggle is an online community for descriptive analysis and predictive modelling. It collects variety of research fields' dataset from data analytic practitioners. Data scientists compete to build the best model for both descriptive and predictive analytic. It however allows individual to access their dataset in order create models and also work with other data scientist to solve various real world analytics problems. The input dataset used in developing this model has been downloaded from Kaggle [23]. The dataset contains design characteristics of iron ore froth flotation processing plant which were put together within three (3) months. This is nicely organized using common format and a standardized set of associate features of iron ore froth flotation system.

5.1.1 Structure of Dataset

The dataset contains 24 columns representing the measurements, 737,453 samples exist. The 24 columns include the date and time of the measurement, which will not be used as an input feature. The last columns of the dataset represent the targets of this prediction task: the percentages of iron ore and silica concentrate, which are highly inversely correlated. Our goal is to predict silica concentrate without the use of iron concentrate. The other 21 columns will be used as features for predicting the target value.

5.1.2 Data Exploration and Pre-processing

Preliminary analysis was carried out to understand the global landscape of the dataset. By summarizing, RStudio generated detail statistics of the dataset such as the mean, median, min and so forth. The aim was to assess skewness of each variable and detecting outliers. It was realized that all variables were stored as numeric with the exception of date variable which was stored as a factor. This however did not pose any challenge since it was further eliminated from the analysis as non-predictive feature. All distribution of variables was examined. A typical example is Figure 2 which shows a snapshot of the distributions of pair variables in the dataset explored.

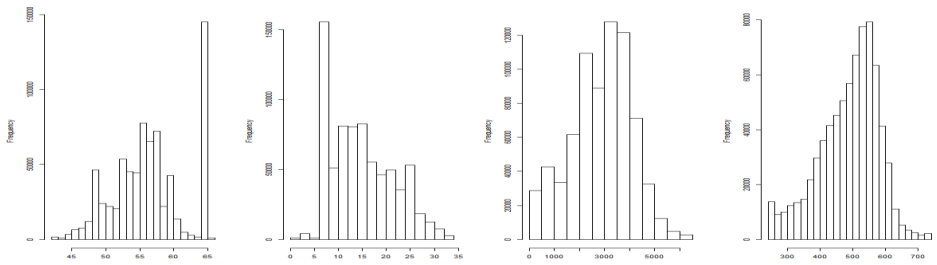


Figure 2 Pair distribution charts for numerical variables-histogram: left-to-right: iron feed, X…silica Feed, Starch Flow and Amina Flow.

As you can see, there is variation in the distribution of each variable as expected. I also established other interesting observations in the outcome variable of interest. The distribution of the silica concentrate in Figure 3 depicts randomness which however buttresses the hypothesis that the froth flotation processing plant is indeed stochastic in nature and difficult to operate.

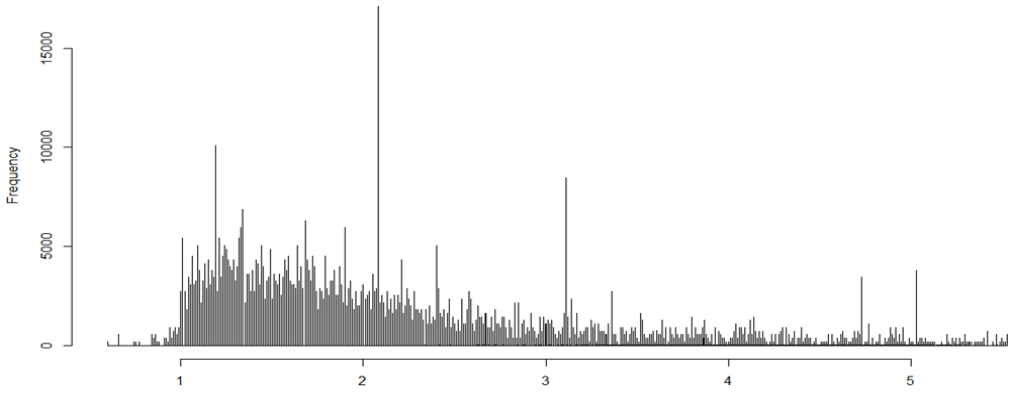


Figure 3 Distribution of Silica concentrate

Moreover, boxplot and threshold adjustment were used to detect outliers and strange numbers. Outliers were observed in Ore-Pulp Flow and Ore-Pulp.pH variables. Excerpt is in Figure 4 which shows the distribution of Ore-Pulp Flow variable's outliers. The outliers were eliminated as can be seen on the same snapshot for further analysis in order to make the analysis trustworthy. A more refined dataset was however, obtained totaling approximately 521188 observations.

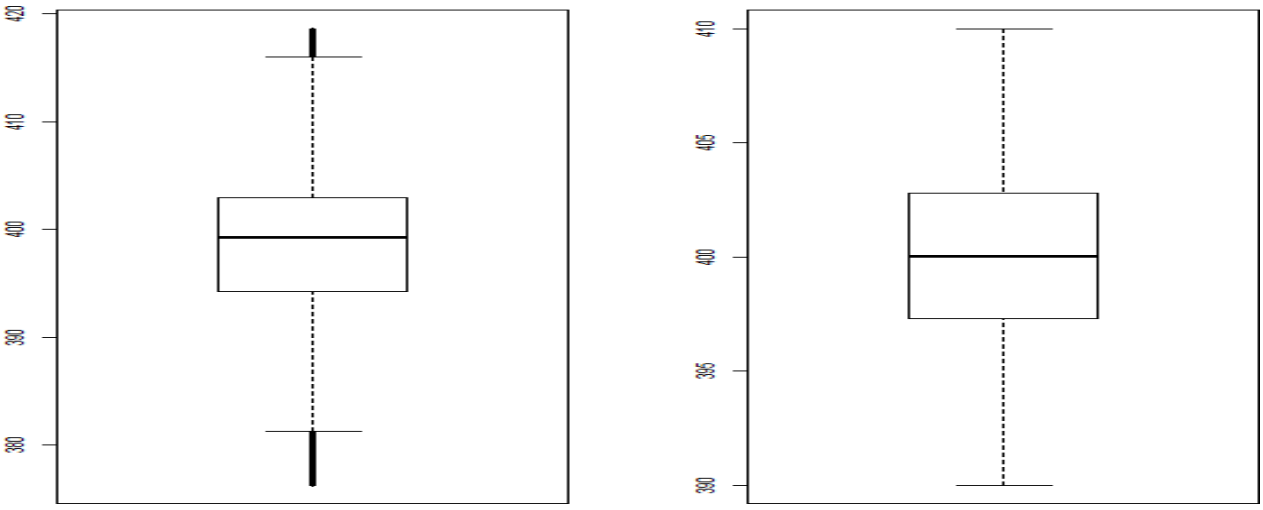


Figure 4 Distribution of Ore-Pulp Flow variable to detect outliers.

Moreover, it is worth mentioning the sampling technique, by which the percentage of silica and iron concentrate was measured in this study. Sample of the slurry known as froth as indicated earlier on is drawn at twelve (12) minutes interlude. The laboratory result usually takes at least 2 hours to be available and only pertains to the amalgamation of such partition. This does not include the process sequence of events at that specific instant. However, lag- transformed was generated on the silica concentrate variable to incorporate vital information that might have lost as the process changes during the 2 hours' dead time.

5.1.3 Correlation Analysis

A correlation analysis was employed in this study to examine if explanatory variables share the same linear relationship with the outcome variable in order to detect duplications of variables in the dataset. Amongst other things, highly correlations between variables were observed in the dataset. The clusters of correlated features were found using Pearson correlation coefficient r cutoff of [-0.5, 0.5]. ThePearson correlation coefficient r, takes a range of values between +1 to -1. A value of 0 indicates that there of no relationship between the two variables. A value less than zero indicate a negativerelationship and a value greater than zero connotes a positive association: that is as one unit of

variable increases, so does the value of the other variable that is if the amount of iron in iron ore increases then % silica in iron ore is decreases.

Using certain tools in python we found quantitatively that % Iron and % Silica concentrate are negatively correlated.

Machine Learning

The steady progress of machine learning has been quite phenomenon over a decade where data is now seen as an important asset to be used judiciously by all companies. Machine learning, in simple term referred to algorithms that learned from data in an iterative manner to identify trends, patterns and correlations. Moreover, it is more applicable especially to datasets where past observations are potent predictor of the future. In this study, among wide variety of existing methods, we employed three supervised machine learning algorithms to a historical dataset observed in the froth floatation plant system.

Supervised machine learning is a process of furnishing the algorithm with observations in which the variable outcome of interest is known beforehand and the algorithm learn from the observed data to make prediction of future values [25].

Modelling

DATA SPLITTING

The dataset (734543,24) is split into train and test dataset at proportional (80%,20%) .Our main aim to find whether we can predict the possibility of prediction of silica without iron concentrate. To achieve these steps:

1. To build the model (with iron concentrate) using train data and predict the silica concentrate with test data and using R^2 as metric we can predict the performance of model
2. To build model (without iron concentrate) using train data and predict the silica concentrate using test data and using R^2 as evaluation metric.
3. Compare two models 1 and model 2 and find out which shows better results.

5.1.4 Selected Methods

Multiple Linear Regressions

Multiple Linear Regression is one of the most popular and simple machine learning algorithms employed in numerical predictive task. It is mainly used to model a relationship between a numerical outcome variable and a set of explanatory variables. In other words, the model is expected to fit a relationship between a numerical outcome variable and a set of predictors. Y (called the response, target, or dependent variable) and a set of predictors X1, X2, X3…….Xn. (also referred to as independent variables, input variables or covariates). The assumption is that the function in Equation 1 approximates the relationship between the predictors and the outcome variable:

Y=$\beta_o + \beta_1 + \beta_1 x_1 + \beta_2 x_2$+.....+$\beta_p \quad x_p + \epsilon$,---------- (1)

$\beta_o \ldots \ldots, \beta_p$ $are\ cofficients$, and $\epsilon$is the noise or unexplained part.

In predictive modeling, data are often used to estimate the coefficient and the unexplained noise that could better lead to prediction of individual records. Regression modeling means not only estimating the coefficients but also choosing which predictor can be included in the model. And the performance of the model is evaluated using a holdout dataset. In R, I use the lm () function to fit the model.

Moreover, Multiple Linear Regression model does exceptionally well with linearly separability phenomena and has a faster computational capability. On the other hand, the limitation is that it does not perform well in non-linearly-separable cases [25].

Random Forest

Random Forest (RF) is an ensemble machine learning proposed by Leo Breiman that combines predictors tree for predictive or classification task based on independent random samples of observations. In order to grow a tree, multiple randomsamples are drawn with replacement from the training dataset. This connotes that each tree would be grown with its version of the training dataset. Moreover, a subset of the explanatoryvariables is also randomly selected at each node during the learning process. After training, prediction scores are however determined by averaging the predictions from all individual regression trees. This helps to attenuate overfitting and variance.

Though, there are many methods that perhaps most likely to have impact on model performance. In this study, I focused on two main tuning hyperparameters that have significant effect on the prediction outcome of Random Forest. I employed repeated cv method to split the dataset into 10 folds cross-validation in [Caret]. For the implementation of the Random Forest in this project, the hyperparameters namely number of trees (*ntree*) and the number of randomly selected features (*mtry*) were experimentally tuned .

- **mtry**- Number of variable is randomly collected to be sampled at each split time.
- **ntree**- Number of trees will grow after each time split

For tuning the hyperparameters, *ntree* and *mtry* of the Random Forest, I used the function"randomForest" available within the R-Package. For tuning *ntree*, I tried values (100, 200, 300, 500, and 1000).
The advantage of using this algorithm is that it does not over fit the data and does exceptionally well with non-linear relationship task. One challenge worth mentioning here is that the algorithm is computationally expensive [26].

DECISION TREE REGRESSOR

Decision Tree utilizes a graph appearing like tree or model of decisions and their possible outcomes containing resource costs and utility. It only incorporates conditional control statements. Every internal node represents a test on a feature. Every edge on the graph becomes the consequence of the related test and every leaf node means a class label or possible value of the item in regression. The model from Scikit-Learn is a 1D regression method with Decision Tree.

ADABOOST

Boosting is an ensemble method which is aimed at forming a strong classifier from a group of comparably weak classifiers. Initially, the model is built from the training data. Afterward, a second model is created by correcting the errors of the first model. Models are incrementally created. AdaBoost regressor is a meta estimator that starts with fitting a regressor on the original dataset and later fits extra versions of the regressor on the same dataset. However, the weights of records are altered according to the current estimation error.

Comparison of models(with iron concentrate and without iron concentrate)

We can predict the % silica with and without iron concentrate and using the above models we have analysed using R^2 METRIC and MEAN-SQUARED-ERROR.The evaluation metric is coefficient of determination (R^2) statistic was used as evaluation metric to compare the performances of the methods. It reflects the possibility of new coming data to be fell within the predicted outcomes.

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

R squared is used to measure statistically the closeness of the tested instance to regression line created by the regression algorithm. It means that it reveals how an unknown data is likely to fall on the regression line. It is the ratio of the explained variation by the model to the total variation. Its value is in the range of [0, 1].

MEAN-SQUARED-ERROR

The Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true value. It is a risk function, corresponding to the expected value of the squared error loss.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

The experimental results show that AdaBoost regressor clearly provided higher coefficient of determination value than other algorithms. This is probably because AdaBoost is an ensemble method, which generally provides better accuracies than an individual model by averaging the decisions of several predictors. In addition, AdaBoost is an iterative algorithm, each time reweighting the instances in the dataset to focus the next classifier on incorrectly classified ones. By this way, it constructs a strong classifier from a combination of weak classifiers. The experimental results demonstrate the superiority of AdaBoost
.

Another ensemble learning algorithm, Random Forest, also obtained a high score 0.96 which is very close to the best value. So, Random Forest can also be used alternatively, especially when there are many input variables, since it randomly selects the subset of the features, which decreases the running time of the algorithm. It can be observed from table that Decision Tree method has also acceptable accuracy.However, ensemble based methods (AdaBoost and Random Forest) have higher accuracy with respect to this method. According to the results given in Table the RIDGE method is not suitable to be the base

learner for multi-output regressor because it performs significantly worse than decision tree based methods.

In this study, a multi-target regression problem is handled to predict quality in a mining process. The aim is to construct a robust model that simultaneously estimates the amount of silica and iron concentrates in the ore. Several approaches are implemented and compared to be able to handle more than one target variable. We tried to observe the performance of a multi target regression approach when target features are highly correlated. At the end, it is noticed that this approach can also be efficient in manufacturing data when a related attribute is not given to the algorithm as an input parameter. Instead, that feature canalso be evaluated as an output variable by being added to the existing target feature. We have observed that this alteration did not create an adverse effect on the regression performance.

6 Performance Test

Performance testing involves measuring the actual time it takes for the machine learning model to make predictions. This is done by recording the start and end times of prediction requests. Here we need to first find the constraints.

During performance testing, we monitored the utilization of hardware resources, such as CPU and memory, while the model is running. This can be done using system monitoring tools.

Scalability testing assesses how the machine learning model handles increasing workloads. You progressively increase the size of the input dataset or the complexity of prediction requests to gauge the model's scaling capabilities.

Response time measurement involves recording the time it takes for the model to respond to a prediction request. This includes processing time and any network latency.

Load testing evaluates how the model performs under expected and peak loads. You gradually increase the number of prediction requests to assess its behavior as load increases.

During performance testing, you analyze the model's resource efficiency, particularly its memory usage and computational requirements. This helps identify areas for optimization.

Performance testing provides valuable insights into how the machine learning model behaves under various operational conditions. By systematically measuring and analyzing these performance aspects, you can ensure that the model functions efficiently and effectively in real-world mining processes,meeting the required performance criteria.

6.1.1 Test Plan/ Test Cases

Objectives: The test plan outlines the objectives of testing, which include verifying the accuracy of the machine learning model in predicting silica and iron concentrations in mining samples.

Scope: It defines the boundaries of testing, specifying that the model will be evaluated using historical mining data for both silica and iron concentration predictions.

Test Environment: The test environment includes the machine learning framework, libraries, and hardware/software requirements for running the tests.

Test Data: Historical mining datasets are used for testing, with information on the data source, dataset size, and date range.

Testing Methodology: The methodology involves selecting a suitable machine learning algorithm, data splitting techniques, and evaluation metrics for regression tasks.

Test Schedule: A timeline is established for testing activities, including data preparation, model training, evaluation, and reporting.

6.1.2 Test Procedure

Preconditions: Ensuring the model is trained and the test environment is set up correctly.

Test Execution: Executing each test case by loading input data, running the model, and comparing predictions to expected values.

Data Validation: Calculating evaluation metrics (e.g., MAE, MSE) to measure prediction accuracy.

Post-Conditions: Recording results, including predicted values and evaluation metrics.

6.1.3 Performance Outcome

Silica Concentration Prediction Performance: Reporting MAE, MSE, RMSE, and R2 values for silica concentration predictions and noting any observed patterns or trends.

Iron Concentration Prediction Performance: Reporting MAE, MSE, RMSE, and R2 values for iron concentration predictions and noting any observed patterns or trends.

Overall Model Performance: Summarizing the overall model performance in predicting both silica and iron concentration and offering insights or recommendations for improvement if needed.

- **Correlation Output of the data:**

The plot of correlations between different Variables in the dataset in shown in figure 5. Each cell's color and possibly thenumeric annotation inside it indicate how strongly and in what direction two Variables are correlated. The heatmap helps to identify patterns and relationships in the data which can be particularly useful in quality predictionin a mining process, as mentioned. It highlights which variables are strongly related to quality and which ones might have a weaker influence, The color level indicates the strongest to weaker influence in the quality ofore.

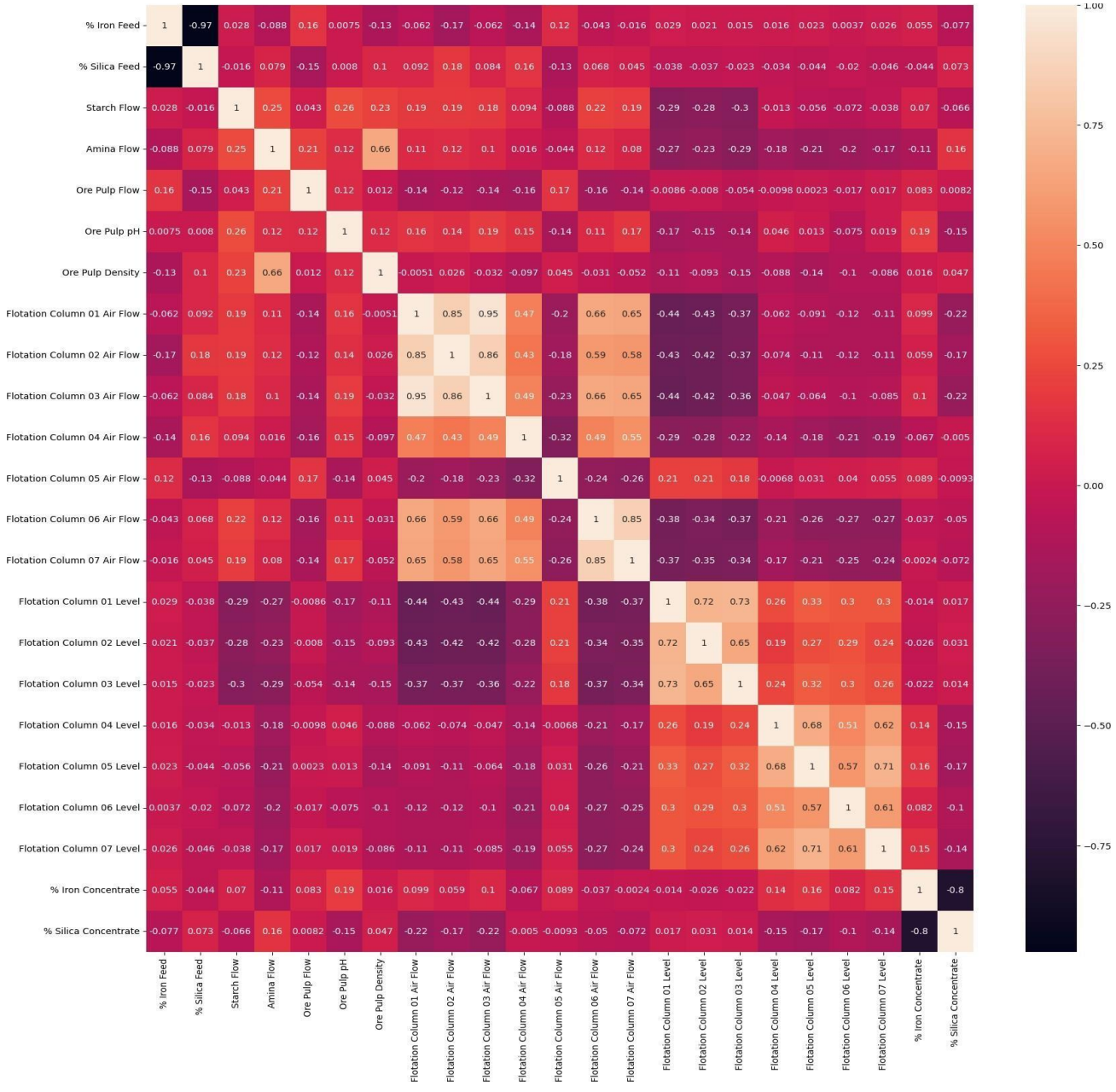


Figure 5: The correlations between different variables in the dataset output.

- **Relation of possible features with %silica:**

A visual exploration of how certain features related to the '% silica concentrate' label using scatter plots is shownin the figure 6 . It helps to understand the potential correlation or patterns between these features and the target label is the context of the quality prediction in a mining process. It contains 8 scatter plots, each plotting a different feature against the '% silica concentrate 'label. Data collection, data preprocessing and data analysis for the project is successfully implemented, further in week [3] planned to develop the trainedmodel and test model for the datasets.

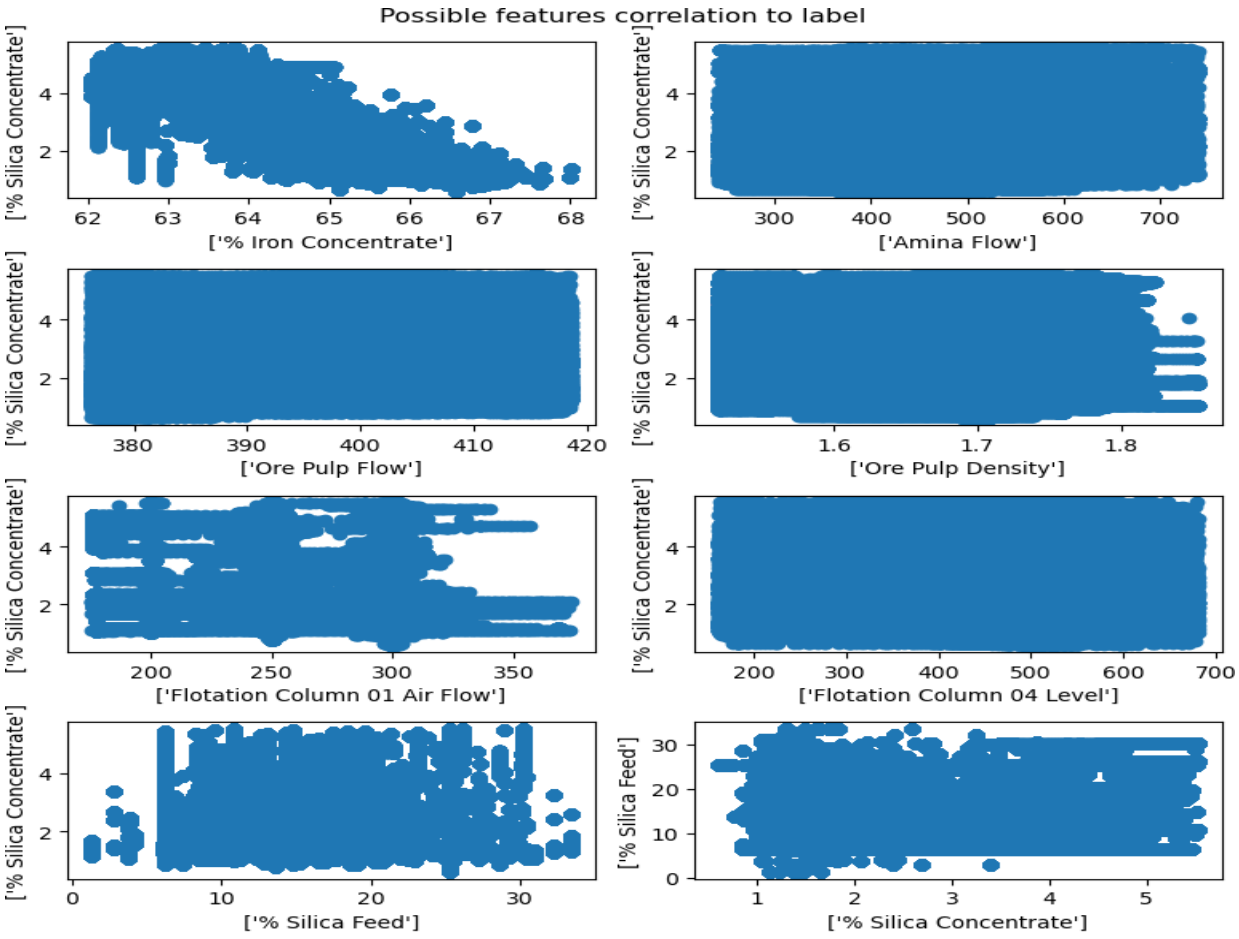


Figure 6: Output of features relates to the '% Silica Concentrate' label using scatterplots

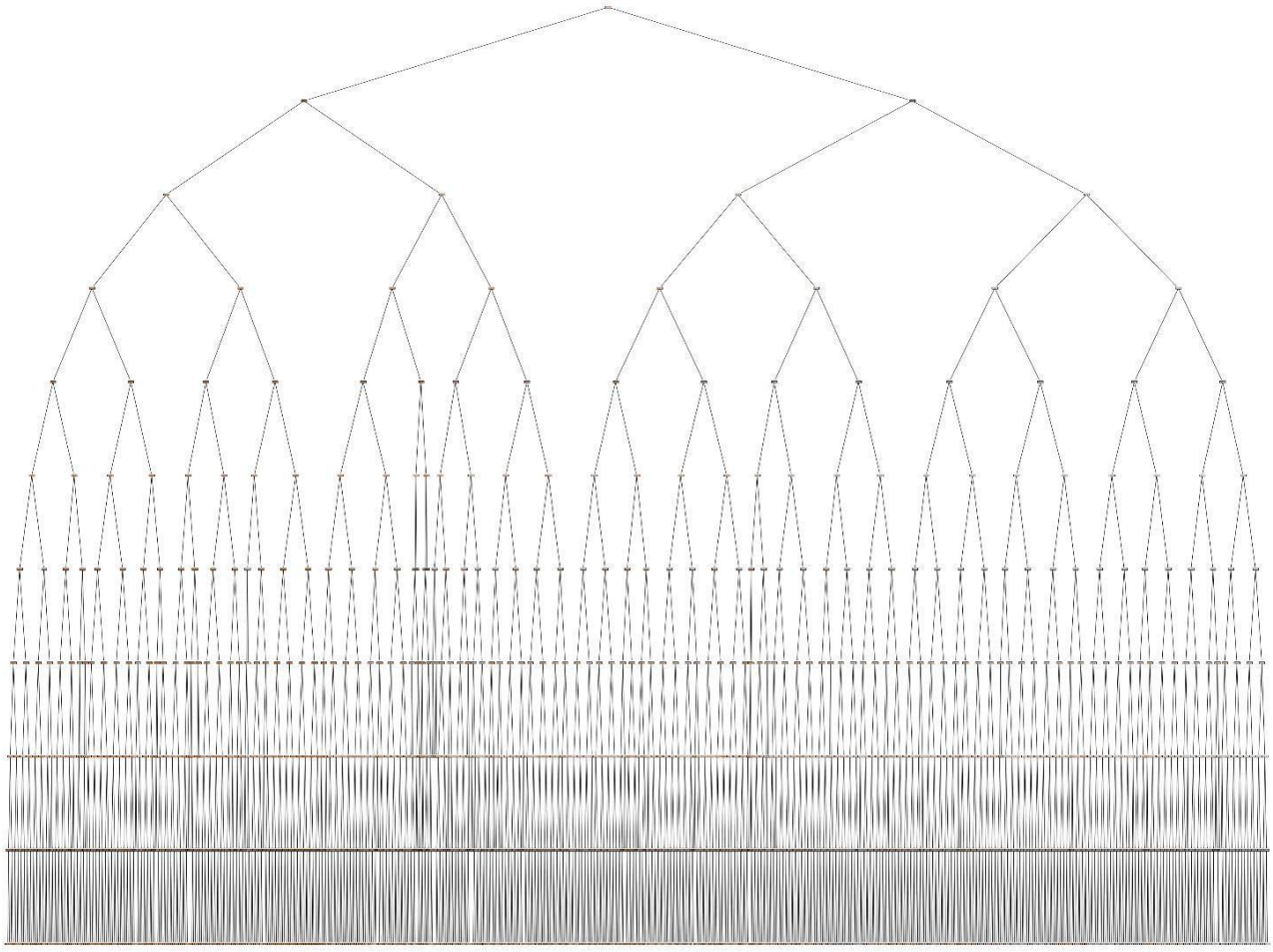


Figure 7 represents the visual of a decision tree of Random Forest Regressorto the normalised training data in the regression model

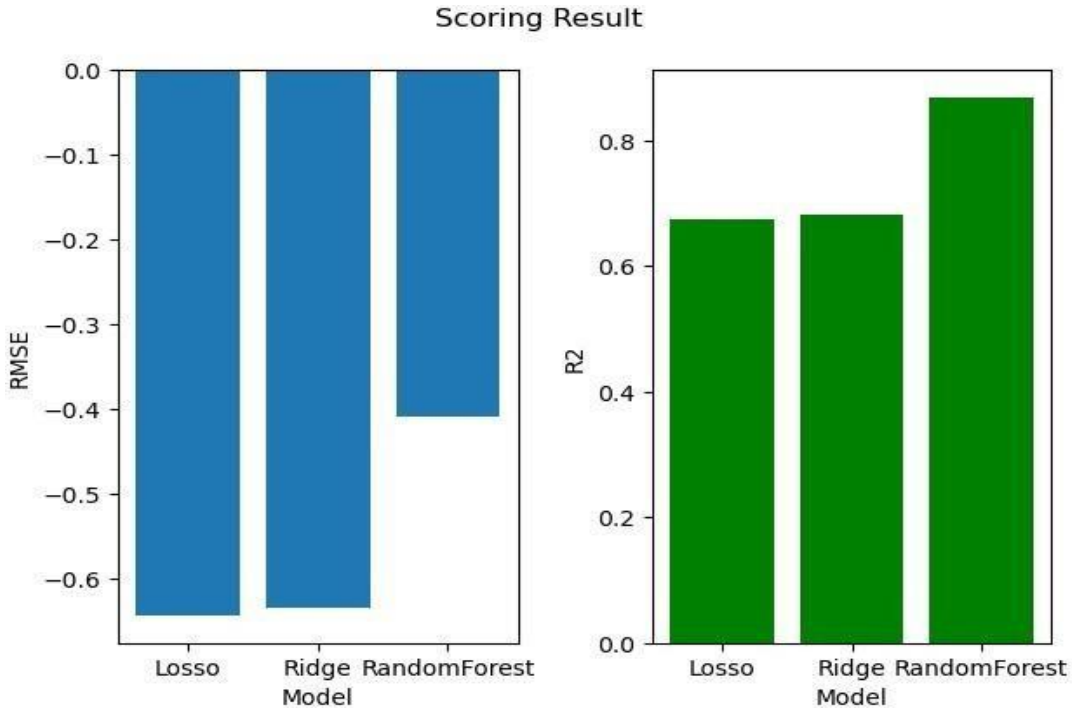


Figure 8: bar plots of scoring results by comparing RMSE and R2

- The subplot1 in figure 8 contains a bar plot (blue in color) where the x-axis represents the different machine learning models (" Losso ", "Ridge", and "Random Forest").
- The y-axis represents the Root Mean Squared Error (RMSE) values associated with each model.
- subplot 2 also contains a bar plot with the same x-axis representing the different machine learning models.
- The y-axis represents the R-squared (R2) values corresponding to each model.
- The X-axis label is set to “Model”, and the Y-axis label is set to “R2”.
- Appearance of the bar plots will depend on the values of RMSE_COL, as well as the specific data associated with each model.
- The graph represents to visualize and compare the RMSE and R2 scores of different machine learning models using bar plots arranged side by side.

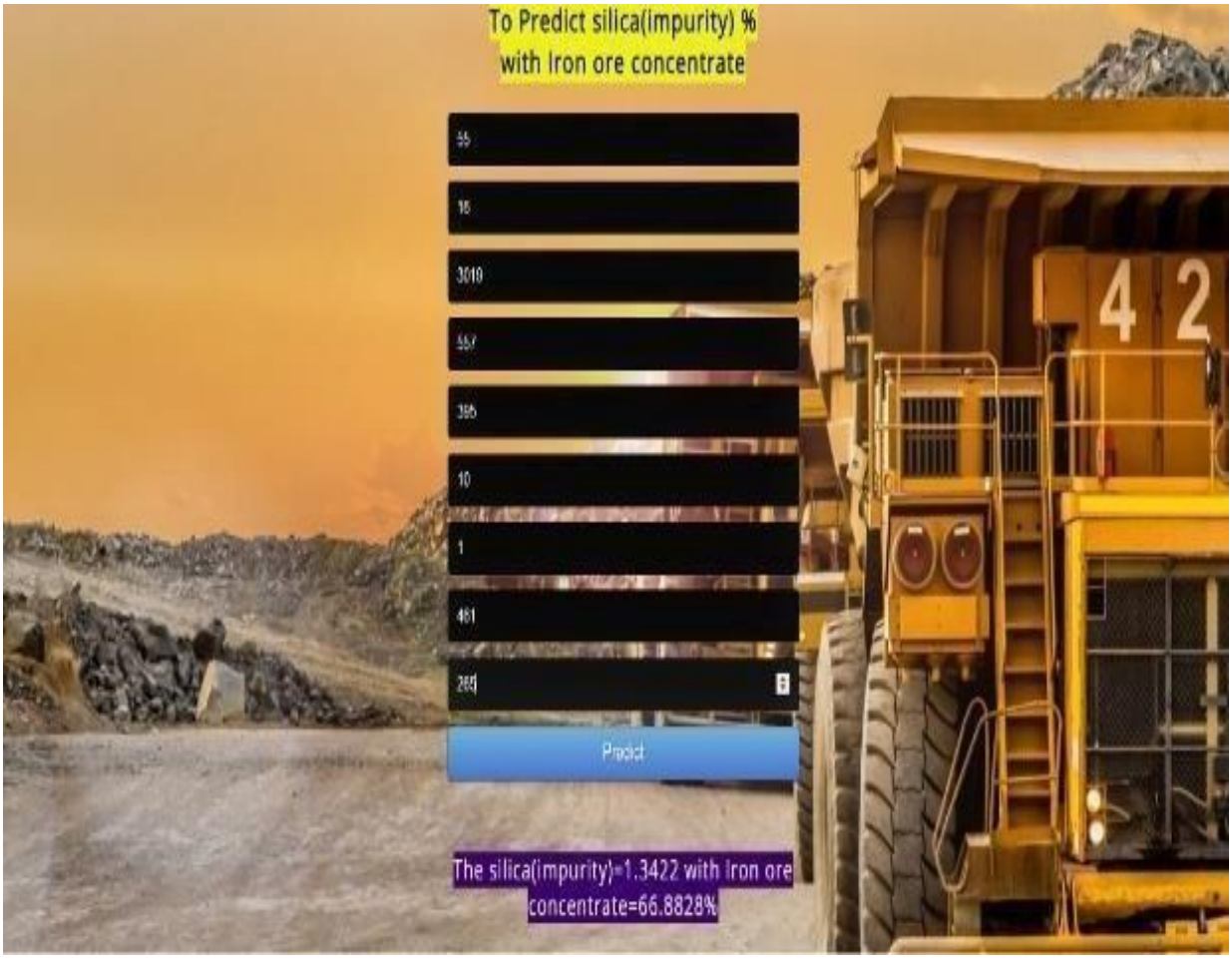


Figure 9: To Predict silica (impurity) % with Iron Ore

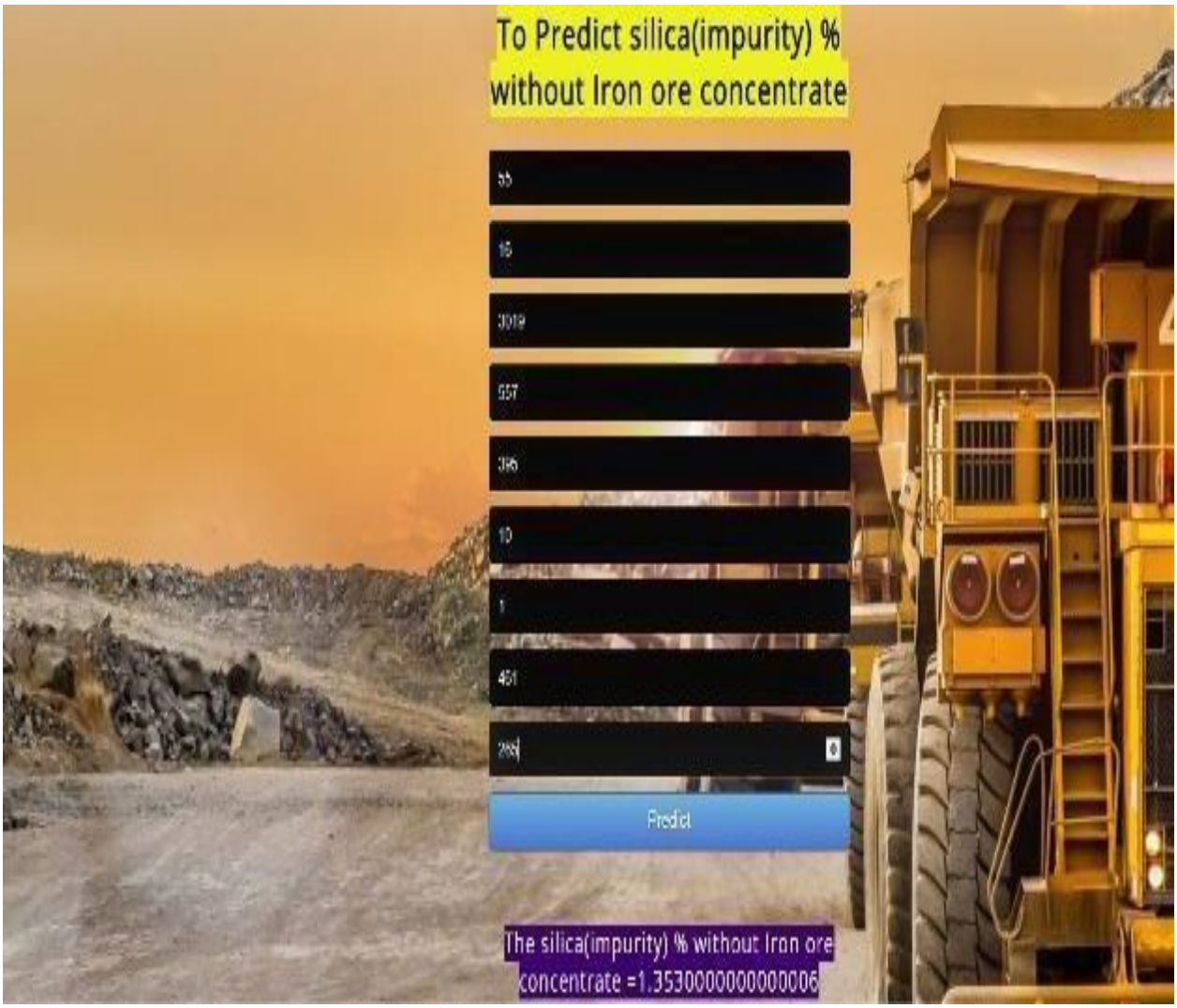


Figure 10: To Predict silica (impurity) % without Iron Ore.

7 My learnings

Brief explanation of the Data Science AI, two career paths sometimes cross one another. Data science uses AI and its many components, including machine learning and deep learning.

It develops and plans the whole big data environment on Hadoop and Spark. Holding experience in data visualization, data mining, and data migration is an essential requirement. Also, to have demonstrated experience with C++, Java, Scala, and Python required to grow. It's crucial to work backward to develop the abilities required for your career, whetherit's in AI or data science.

Artificial Intelligence method works by consuming huge amounts of labeled training data. Then analyzing the data for correlations and patterns. AI uses these patterns to make predictions about business decisions. The replication of human intelligence by computer systems is defined as artificial intelligence. Speech recognition, natural language processing, and machine vision are some good examples of AI applications. Data Science is a broad process that includes pre-treatment, analysis, and visualization.

Then it comprises prediction and generating understanding. It is the study of data to pull valuable understanding for business. It combines various disciplines to evaluate huge data.

8 Future work scope

In this study, different methods of supervised machine learning predictive model performance would be evaluated using froth flotation processing plant dataset. R statistical capacities for modeling would be applied to the data and therefore no attempt is being made to propound any novel predictive algorithm. However, froth physical characteristics such as bubble shape, speed, size distribution and colors were not considered as model input parameters.

In this study, a multi-target regression problem is handled to predict quality in a mining process. The aim is to construct a robust model that simultaneously estimates the amount of silica and iron concentrates in the ore. Several approaches are implemented and compared to be able to handle morethan one target variable. We tried to observe the performance of a multi target regression approach when target features are highly correlated. At the end, it is noticed that this approach can also be efficient in manufacturing data when a related attribute is not given to the algorithm as an input parameter. Instead, that feature can also be evaluated as an output variable by being added to the existing target feature. We have observed that this alteration did not create an adverse effect on the regression performance.

We can implement LSTM as DL model and predicted the results we can also use CNN+LSTM to predict the results