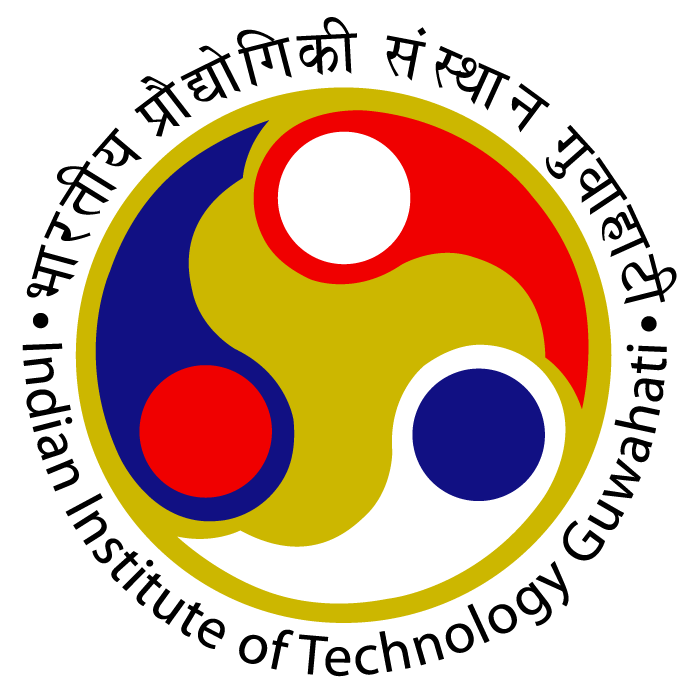
Quality Prediction in a Mining Process

**Shashank Kumar Gupta**

**210107077**

**Submission Date: August 21, 2024**



**Final Project submission**

**Course Name : Applications of Al and ML in chemical engineering**

**Course Code: CL653**

Contents

[1 Executive Summary 3](#_Toc163320085)

[2 Introduction 3](#_Toc163320086)

[3 Methodology 3](#_Toc163320087)

[4 Implementation Plan 5](#_Toc163320088)

[5 Testing and Deployment 7](#_Toc163320089)

[6 Results and Discussion 8](#_Toc163320090)

[7 Conclusion and Future Work 11](#_Toc163320091)

[8 References 11](#_Toc163320092)

[10 Auxiliaries 11](#_Toc163320094)

# Executive Summary

# The project addresses the challenge of predicting the percentage of silica concentrate in a mining process, aiming to empower engineers with early information to optimize ore quality and reduce environmental impact. The proposed solution involves leveraging machine learning techniques on a dataset comprising process variables and quality measures. Methodologies include data preprocessing, feature selection, and building predictive models using Linear Regression (Lasso and Ridge models) and Random Forest Regressor algorithms. The expected outcome is the development of an accurate predictive model that forecasts silica concentration, enabling proactive decision-making to enhance operational efficiency and environmental sustainability.

# Introduction

**Background:** In the field of Mining, quality control of mining processes is of paramount importance. The efficiency of these processes directly impacts both economic profitability and environmental sustainability. The presence of impurities, such as silica, in ore concentrate can significantly affect downstream processing and product quality. Therefore, accurately predicting and controlling the percentage of silica concentrate is crucial for optimizing process efficiency and minimizing environmental impact.

**Problem Statement:** The main problem addressed in this project is the prediction of the percentage of silica concentrate in a mining process for enhanced operational efficiency, reduced environmental impact, and optimal resource utilisation.

**Objective:** Develop a machine learning model capable of predicting the percentage of silica concentrate in iron ore pulp.

# Methodology

**Data Source:** The data is collected from a real time mining plant of iron ore and it comprises comprehensive information on the froth flotation process in mining operations. The dataset provides a detailed record of quality measures, process variables, and ore characteristics collected over a specific time period. It includes information such as timestamps, quality measures of the iron ore pulp, process variables related to flotation columns, and final iron ore pulp quality measurements from laboratory analysis. There is total 24 columns of data, including different columns levels as well as air flow in those columns.

**Data Preprocessing:** To ensure the dataset is clean and ready for analysis, the first step would be to address missing values by either deleting them or imputing them depending on their prevalence. Next, categorical variables would be properly encoded, and numerical features would be normalized to a consistent scale. Normalization techniques like Z-score normalization or Min-Max scaling would be employed to ensure all features have comparable ranges and distributions. This step is crucial for preventing larger magnitude features from dominating the model training process. Lastly, the dataset would be divided into training, validation, and testing sets to facilitate model evaluation and ensure robustness. By following these steps, the dataset would be prepared for effective machine learning analysis, enhancing predictive power and model performance.

**Model Architecture:** Model architecture comprises three main models: Random Forest, Linear Regression with Lasso regularization, and Linear Regression with Ridge regularization. Random Forest is chosen due to its ability to handle complex interactions and heterogeneous data, making it suitable for capturing nonlinear correlations between process factors and silica concentration in the ore concentrate. Its ensemble approach reduces overfitting and enhances generalization capabilities, while also offering insights into feature significance, critical for identifying influential factors in silica concentration estimation. Linear Regression with Lasso regularization is selected for its ability to produce sparse models by penalizing the absolute values of coefficients, effectively performing feature selection and simplifying the model. This is advantageous when dealing with a large number of features, enhancing model interpretability and prediction performance by focusing on the most relevant predictors of silica concentration. Similarly, Linear Regression with Ridge regularization is chosen to manage potential multicollinearity and reduce model complexity by penalizing large coefficients. This improves model stability and prevents overfitting, especially with noisy or redundant features. Ridge Regression provides interpretable coefficients, aiding in understanding the impact of different variables on silica concentration prediction and delivering accurate and reliable forecasts.

**Tools and Technologies:** The project is supported by essential libraries such as NumPy and Pandas for data manipulation. Visualization tasks are facilitated by Seaborn and Matplotlib, while Scikit-learn is utilized for machine learning implementations, including model selection and evaluation. Additionally, the multiprocessing library enables efficient parallel computing tasks

# Implementation Plan

**Development Phases:** I have divided the complete Model development procedure into 3 phases –

**1st Phase: Data Preprocessing and EDA analysis of Data (12th April)**

During the initial phase of the project the focus was on data preprocessing and exploration. This involves locating and addressing missing values either through deletion or imputation methods. Categorical variables will be encoded, and numerical features normalized using techniques such as Z-score normalization or Min-Max scaling to ensure consistency across variables. Exploratory data analysis (EDA) will be done to understand the distribution of variables and uncover potential patterns or correlations within the dataset. Furthermore, feature selection techniques will be applied to identify the most relevant predictors of silica concentration, laying the groundwork for subsequent model development stages.

**2nd Phase: Model Training and Hypertuning (14th April)**

We would be using three sets of models and hypertuned them –

1. **Linear Regression Model ( Lasso Model ):** Lasso applies L1 regularization, which can lead to sparser models by penalizing the absolute values of the coefficients. This encourages some coefficients to become zero, effectively performing feature selection and simplifying the model. By reducing the impact of less important variables, Lasso helps improve model interpretability and potentially enhances prediction performance by focusing on the most relevant features. This can be particularly beneficial when dealing with a large number of features, as it helps in identifying and retaining only the most significant predictors of silica concentration, leading to more efficient and accurate predictions.
2. **Linear Regression Model ( Ridge Model ):** Ridge Regression helps manage potential multicollinearity and reduces model complexity by applying L2 regularization, which penalizes large coefficients. This leads to improved model stability and prevents overfitting, especially with a dataset that may contain noisy or redundant features. Ridge Regression can effectively handle a range of continuous input features, such as those in your dataset, and provide interpretable coefficients, allowing you to understand the impact of different variables on the predicted silica concentration. By balancing the model's bias-variance trade-off, Ridge Regression can deliver accurate and reliable predictions, aiding engineers in making informed decisions based on the predicted impurity levels in the ore concentrate.
3. **Random Forest ( Regressor Model ):** Because Random Forest can handle complex interactions and heterogeneous data, it is one of the AI/ML models being considered for this project. Because Random Forest can capture nonlinear correlations between process factors and silica content in the ore concentrate, it is especially well-suited for this purpose. Random Forest's ensemble approach, which blends several decision trees trained on bootstrapped subsets of the data, reduces overfitting and enhances generalisation capabilities. Furthermore, because Random Forest naturally offers insights on feature significance, it is possible to identify the critical factors influencing estimates of silica concentration. Because Random Forest can handle a variety of data types, including both numerical and categorical features, it is a good contender to properly estimate silica concentration given the dataset's diversified nature.

In Linear Regression Models, hyperparameter tuning is performed using grid search with cross-validation (‘GridSearchCV’). The hyperparameter being tuned is the regularization parameter alpha for Lasso regression. The range of alpha values to try is defined using ‘np.logspace(-4, 0, 100)’. Grid search iterates over these alpha values to find the one that minimizes the mean squared error. The best alpha value is then used to train the final Lasso regression model. This process helps to optimize the model's performance by selecting the best regularization strength.

In Random Forest Model, hyperparameter tuning is conducted using grid search with cross-validation (`GridSearchCV`). The parameters being tuned are `max\_depth` (representing the maximum depth of the trees) and `n\_estimators` (representing the number of trees in the forest). The grid search iterates over different combinations of these parameters to find the combination that minimizes the mean squared error. The best parameters are then used to train the final Random Forest Regressor model. This process optimizes the model's performance by selecting the most suitable values for its key hyperparameters.

**3rd Phase – Model Evaluation (25th April)**

For model evalution, I have use RMSE and R2 score. The smaller the value of RMSE for a model, the better the Prediction of the Model.

# Testing and Deployment

**Testing:** The model will be tested against unseen data using a robust testing strategy. This strategy involves splitting the available dataset into training and testing sets, where the training set is used to train the model with for hyperparameter tuning, ensuring generalizability. Subsequently, the tuned model is evaluated on the separate testing set to assess its performance metrics such as RMSE and R², providing insights into its effectiveness in making predictions on new, unseen data. Additionally, visual analysis, like scatter plots comparing actual versus predicted values, aids in understanding the model's predictive behaviour and potential areas of improvement.

**Values of RMSE for Different Models:**

Linear Regression Model (Lasso Model): 0.6384

Linear Regression Model (Ridge Model): 0.6308

Random Forest Regressor Model: 0.2298

**Value of R2 for Different Models:**

Linear Regression Model (Lasso Model): 0.6709

Linear Regression Model (Ridge Model): 0.6787

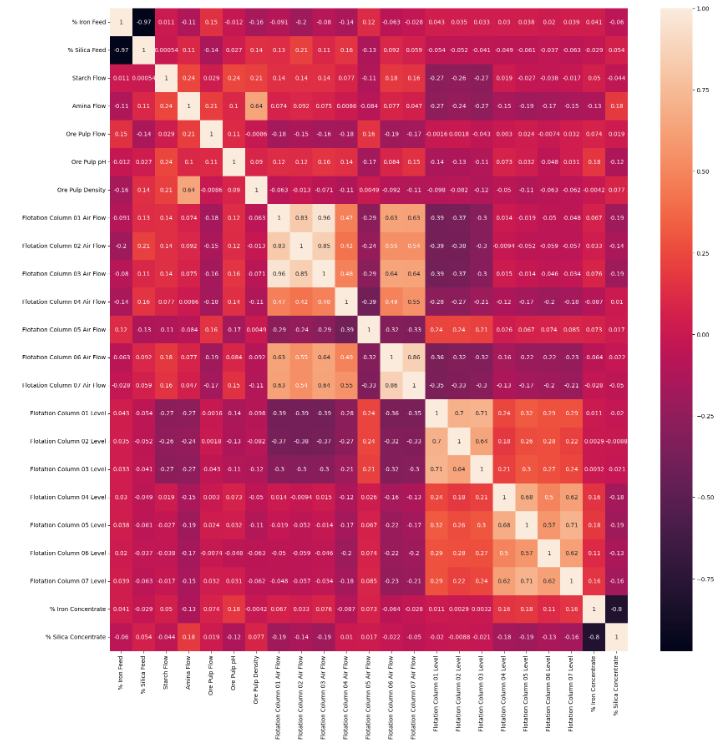
Random Forest Regressor Model: 0.9574

**Deployment Strategy:** Right now, deployment hasn't been implemented, but there's a possibility of doing so in the future, leveraging platforms such as Docker to facilitate seamless scalability and deployment across different environments. Ensuring optimal performance will be achieved through continuous monitoring and updates, while automated pipelines will streamline the deployment process.

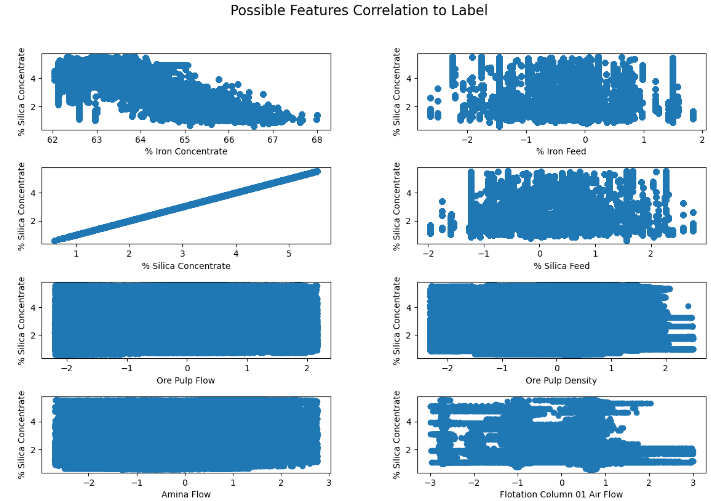
**Ethical Considerations:** The act of deploying the model brings forth ethical considerations concerning fairness, privacy, and transparency. To uphold fairness, it's essential to address biases in the data and monitor model predictions to prevent discriminatory outcomes. Safeguards for privacy should be in place to protect sensitive information, while providing transparent documentation and explanations of the model to users to foster trust and accountability. Regular audits and engaging stakeholders will help tackle emerging ethical challenges throughout the deployment process.

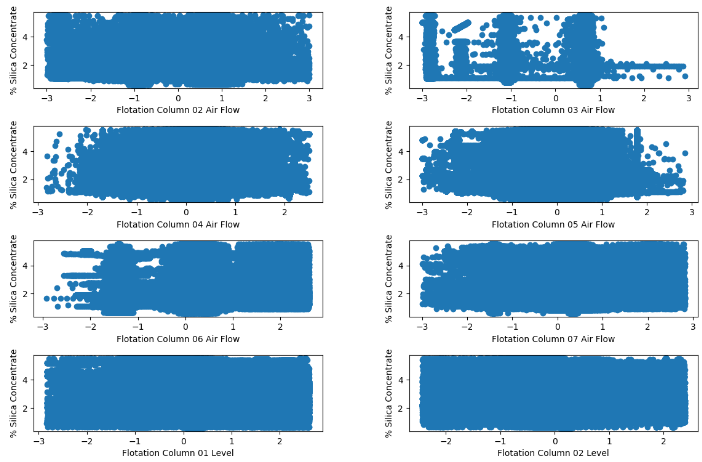
# Results and Discussion

**Heatmap Between Different Features:**

****

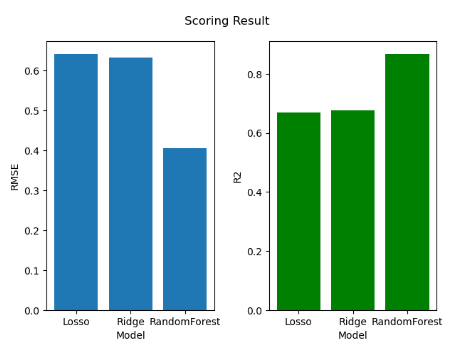
**Graph Between Different Columns and % Sillica Concentrate:**

****

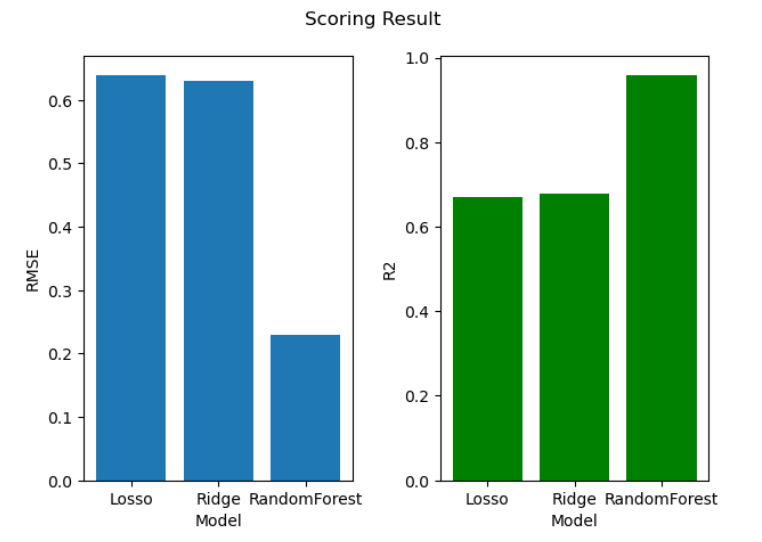
****

It can be observed that, Random Forest Regressor Model performed exceptionally good after hyperparameter tuning, this can be attributed to the optimal selection of hyperparameters. These parameters enable the model to better capture the underlying patterns in the data, resulting in improved classification accuracy and generalization ability.

**Comparative Analysis:**

Before hyperparameter tuning,

After Hyperparameter Tuning,



# Conclusion and Future Work

The project focuses on predicting the percentage of silica concentrate in a mining process, aiming to empower engineers with early information to optimize ore quality and reduce environmental impact. By employing machine learning techniques such as Random Forest, Linear Regression with Lasso regularization, and Linear Regression with Ridge regularization, the project seeks to develop accurate predictive models capable of forecasting silica concentration. The successful implementation of these models could significantly impact the mining industry by enhancing operational efficiency, reducing waste, and improving environmental sustainability. The RMSE and R2 score was less for this model because I dropped the “Date” Column from the data frame.

Future research directions may include exploring advanced machine learning algorithms, incorporating additional process variables, and leveraging real-time data and do not dropping the data column for dynamic predictive modeling to further enhance the predictive capabilities of the models and address evolving challenges in mining processes.

# References

* https://www.kaggle.com/datasets/edumagalhaes/quality-prediction-in-a-mining-process/data

# Auxiliaries

**Data Source:** https://www.kaggle.com/datasets/edumagalhaes/quality-prediction-in-a-mining-process/data