Industrial Internship Report on

”Quality Prediction in Data mining process”

Prepared by

Dhruv Sharma

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 project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time.

My project was **Quality Prediction in Data mining process.**

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 3
2 Introduction 5
2.1 About UniConverge Technologies Pvt Ltd 5
2.2 About upskill Campus....9
2.3 Objective 11
2.4 Reference 11
2.5 Glossary....**Error! Bookmark not defined.**
3 Problem Statement....12
4 Existing and Proposed solution 15
5 Proposed Design/ Model 16
5.1 High Level Diagram (if applicable) 17
5.2 Low Level Diagram (if applicable) 18
5.3 Interfaces (if applicable)....19
6 Performance Test 19
6.1 Test Plan/ Test Cases 19
6.2 Test Procedure....21
6.3 Performance Outcome....28
7 My learnings....29
8 Future work scope 30

1 Preface

Summary of the whole 6 weeks' work.

1. Identification of the pertinent data sources for the mining process, such as geological surveys, drilling activities, and ore processing, is step one in the data collection process.

a. Gather historical data from these sources, making sure it includes a variety of mining activities and spans a sizable time period.

b. Carry out data validation and cleansing procedures to ensure data accuracy and integrity.

2. Data cleaning should be done to deal with missing values, outliers, and discrepancies.

a. Use feature selection or engineering to find the characteristics that are most important for predicting quality.

b. Standardize or normalize the data to provide uniform scaling across various attributes.

3. Exploratory Data Analysis (EDA):

a. Analyze the data to find patterns and relationships using descriptive statistics and visualizations.

b. Recognize how input variables and the goal variable (quality parameters) are related.

4. Model selection:

a. Based on the attributes of the data and the objectives of the project, select appropriate machine learning methods.

b. To get the best model, compare various methods using the right criteria.

5. Data must be divided into training and testing sets for model development.

a. Put the chosen algorithms into practice while fine-tuning their hyperparameters using grid search or cross-validation.

b. Educate the models to correctly forecast quality parameters.

6. Model evaluation:

a. Use the testing dataset to evaluate the model's performance.

b. Examine metrics to assess precision and generalization, such as mean squared error or R-squared.

c. Decide which model for quality prediction performs the best.

7. Deploy the finished model for batch or real-time prediction during the mining process.

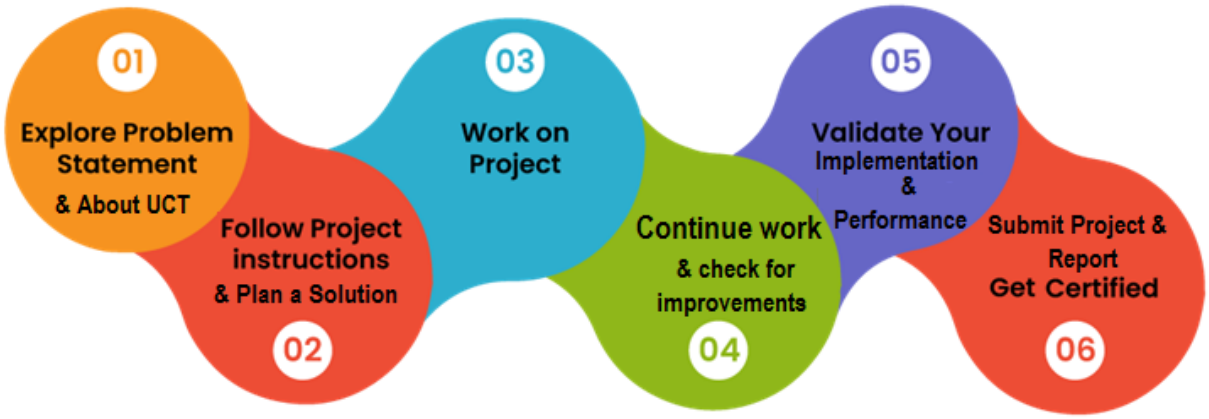
a. Keep an eye on model performance and make any necessary modifications.

8. Continuous Improvement:

a. Regularly gather new data to update and improve the model.

b. Track forecasts and assess discrepancies by contrasting them with actual measurements.

c. Regularly update the model with fresh information or new methods.



Your Learnings and overall experience.

Thank to all (with names), who have helped you directly or indirectly.

Your message to your juniors and peers.

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.

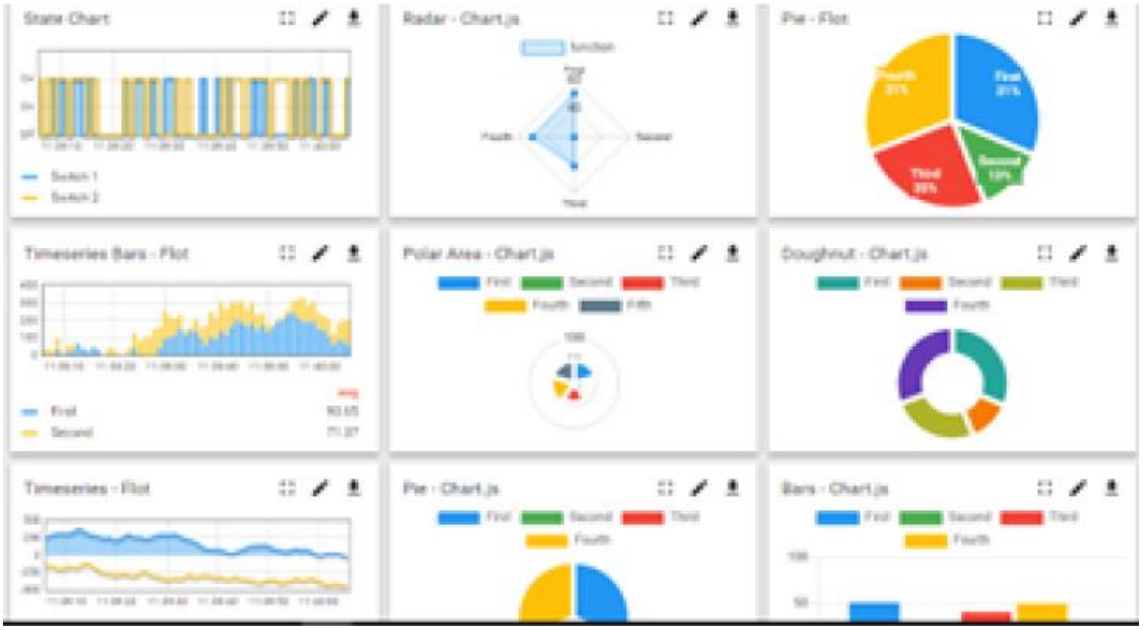
i. UCT IoT Platform (UCT Insight)

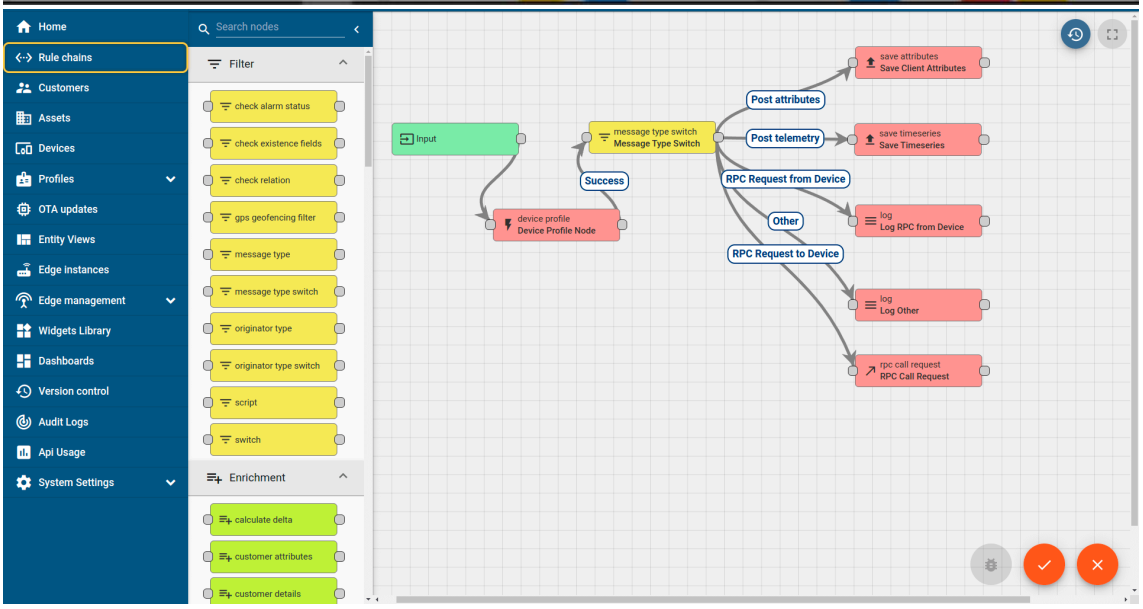
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- 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.

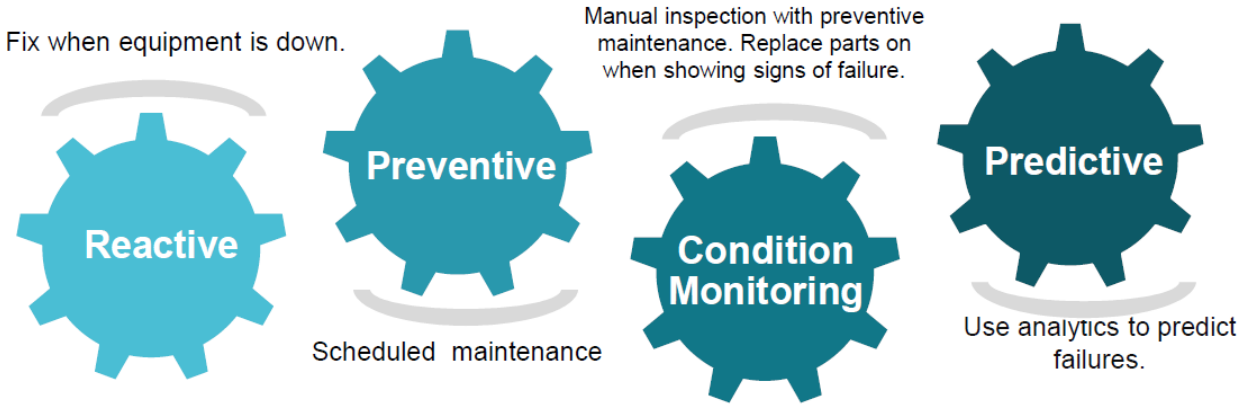
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1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

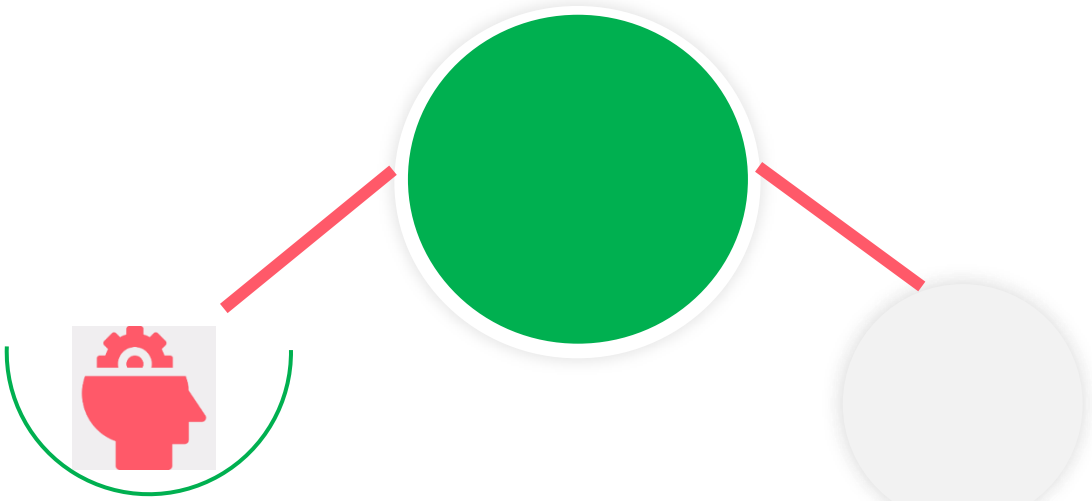
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



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.

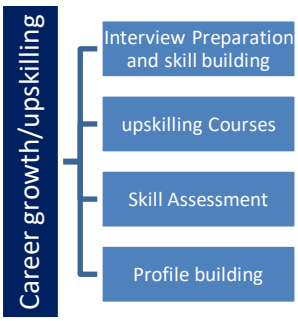
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.

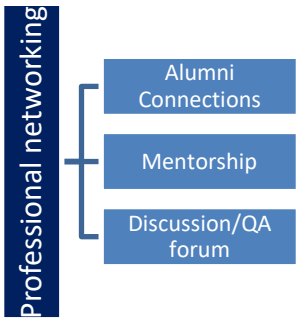


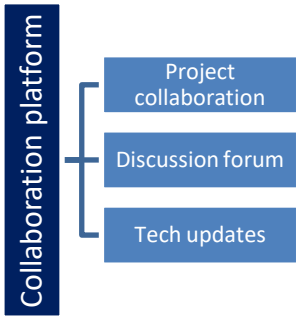
Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

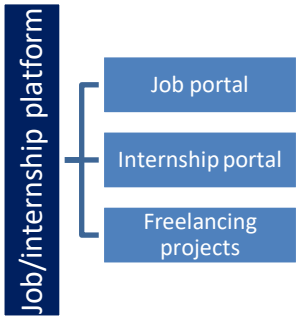
upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/









2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

To Learn Data Science and machine learning and the intricacies involving them

Along with a working real-world problem project.

2.5 Reference

- SciELO - Brazil - Simultaneous use of direct and reverse flotation in the production of iron ore concentrate plant Simultaneous use of direct and reverse flotation in the production of iron ore concentrate plant
- https://ieeexplore.ieee.org/abstract/document/8907120
- https://www.kaggle.com/ankitjha/comparing-regression-models
- .https://www.kaggle.com/rogerbellavista/randomforestregressor-mae-0-0922-rmse-0-2314
- https://www.kaggle.com/plbescond/quality-prediction-r-0-81-mse-0-12
- https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74
- https://towardsdatascience.com/end-to-end-time-series-analysis-and-modelling-8c34f09a3014
- https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779
- https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

3 Problem Statement

Iron can be economically recovered from iron ores, which are a plentiful source of metallic iron. These ores can be found as magnetite, hematite, goethite, limonite, or siderite and come in a variety of hues, including dark grey, bright yellow, deep purple, or rusty red.

Ores with significant amounts of hematite or magnetite, often greater than 60% iron, are referred to as "natural ore" or "direct shipping ore." Typically, 3-7% silica impurities can be found in magnetite iron ore concentrates.

SILICA CONCENTRATE PREDICTION'S IMPORTANCE

Predicting the presence of silica, which is regarded as an impurity in iron ore, aids engineers in the early phases of production. Impurities may be correctly predicted, which will lower the amount of ore delivered to tailings and benefit the environment.

Determining the silica concentration in the ore concentrate is essential since high silica content can also lead to greater slag volumes.

USING DEEP LEARNING AND MACHINE LEARNING

The silica concentration can be accurately predicted using machine learning (ML) techniques. To cut operational costs and deliver quicker forecasts, ML models can take the place of time-consuming and expensive chemical analysis.

"A machine learning breakthrough would be worth ten Microsofts." (Bill Gates)

SOURCES OF DATA

A trustworthy dataset is essential for effective machine learning. Kaggle provided the dataset for this study, "Quality Prediction in a Mining Process," which was used in the analysis.
https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process

DATE OVERVIEW

The dataset has 24 attributes and 737,453 data points. The latter two properties (% iron and % silica concentrate) were employed as goal variables in this study, whereas the previous 21 attributes were used as independent factors.

```
[9]  data.info()

...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 737453 entries, 0 to 737452
     Data columns (total 24 columns):
      #   Column                        Non-Null Count   Dtype
     ---  ------                        --------------   -----
      0   date                          737453 non-null  object
      1   % Iron Feed                   737453 non-null  object
      2   % Silica Feed                 737453 non-null  object
      3   Starch Flow                   737453 non-null  object
      4   Amina Flow                    737453 non-null  object
      5   Ore Pulp Flow                 737453 non-null  object
      6   Ore Pulp pH                   737453 non-null  object
      7   Ore Pulp Density              737453 non-null  object
      8   Flotation Column 01 Air Flow  737453 non-null  object
      9   Flotation Column 02 Air Flow  737453 non-null  object
      10  Flotation Column 03 Air Flow  737453 non-null  object
      11  Flotation Column 04 Air Flow  737453 non-null  object
      12  Flotation Column 05 Air Flow  737453 non-null  object
      13  Flotation Column 06 Air Flow  737453 non-null  object
      14  Flotation Column 07 Air Flow  737453 non-null  object
      15  Flotation Column 01 Level     737453 non-null  object
      16  Flotation Column 02 Level     737453 non-null  object
      17  Flotation Column 03 Level     737453 non-null  object
      18  Flotation Column 04 Level     737453 non-null  object
      19  Flotation Column 05 Level     737453 non-null  object
     ...
      22  % Iron Concentrate            737453 non-null  object
      23  % Silica Concentrate          737453 non-null  object
     dtypes: object(24)
     memory usage: 135.0+ MB
```

Figure 1: Columns in a Dataframe

Process variables are unrelated variables that are utilized to forecast results.

Target variables: Multiple continuous variables that can all be predicted simultaneously when a common set of properties is provided.

MAIN OBJECTIVES

Our concern is whether

- % Iron Concentrate is correlated with % Silica Concentrate
- predict the % silica concentrate without using % iron concentrate .
- If it is correlated and we can predict both % Iron and Silica concentrate at same time using power of ML and DL .

4 Existing and Proposed solution

Existing Solutions given without proper comparison for performance among different applicable ML algorithms

Proposed solution to do a proper comparison among all suitable models and then choosing the one with best performance.

4.1 Code submission (Github link)

https://github.com/DhruvS278/Quality_predicition_in_mining

4.2 Report submission (Github link) : first make placeholder, copy the link.

https://github.com/DhruvS278/Quality_predicition_in_mining

Proposed Design/ Model

Introduction to MULTI-TARGET REGRESSION

Regression models with numerous dependent variables are referred to as multi-target models. In these circumstances, a multi-output regressor is used to discover the relationship between the input characteristics and the aggregate output variables. In this work, we use the multi-target regression approach to estimate the amounts of silica and iron concentrates in the ore and forecast the quality of a mining process.

To do this, we combine two related single-target regression problems into a single multi-output regression job. The iron concentrate has been used as an input parameter in previous models, either with or without trying to predict the silica concentrate. However, the focus of our work is on simultaneously calculating the output variables for both iron and silica concentrations. We evaluate the performance of several multi-target regression methods using measures like the coefficient of determination (R2) and mean squared error (MSE), as well as Random Forest, AdaBoost, XGBOOST, RIDGE, and Decision Tree.

WAYS OF IMPLEMENTING

Methods for issue transformation: These techniques divide a multi-output regression problem into a few single-target problems. Each target variable is modeled separately, and concatenated predictions are used. These techniques, however, fail to consider the connections between the targets, leading to independently predicted targets that can lower the accuracy of the entire forecast.

Algorithm adaptation techniques: These techniques expand and change current algorithms for single-output issues, such support vector machines, to directly handle multi-output datasets. These techniques are more difficult since they try to anticipate the values of several targets while interpreting their interdependence.

METHODS OF MULTI-TARGET REGRESSION

In contrast to earlier research, our goal is to create a multi-target regression model for the mining industry utilizing quality measurements gathered throughout the manufacturing process. We make use of the iron and silica concentrations in the ore, two strongly associated output properties. We use several algorithms as basis learners for the multi-output regressor approach offered by the Python Scikit Learn ML Library through a few experimental experiments on the same dataset.

WHY MULTI-TARGET REGRESSION IS IMPORTANT

Given their substantial association, our research makes a significant contribution by jointly computing these two variables.

MODELS OF MULTI-TARGET REGRESSION

We show the performance of the models and a summary of the techniques employed.

Regression machine learning techniques that naturally allow multiple outputs include some of the following. The scikit-learn library's prominent algorithms including LinearRegression, KNeighborsRegressor, DecisionTreeRegressor, and RandomForestRegressor are among many that fall within this category. To assess their performance, we test both naturally occurring and artificially occurring multi-output algorithms.

We employ models like Ridge Regression, XGBoost, Decision Tree, AdaBoost, and RandomForest, as was previously described.

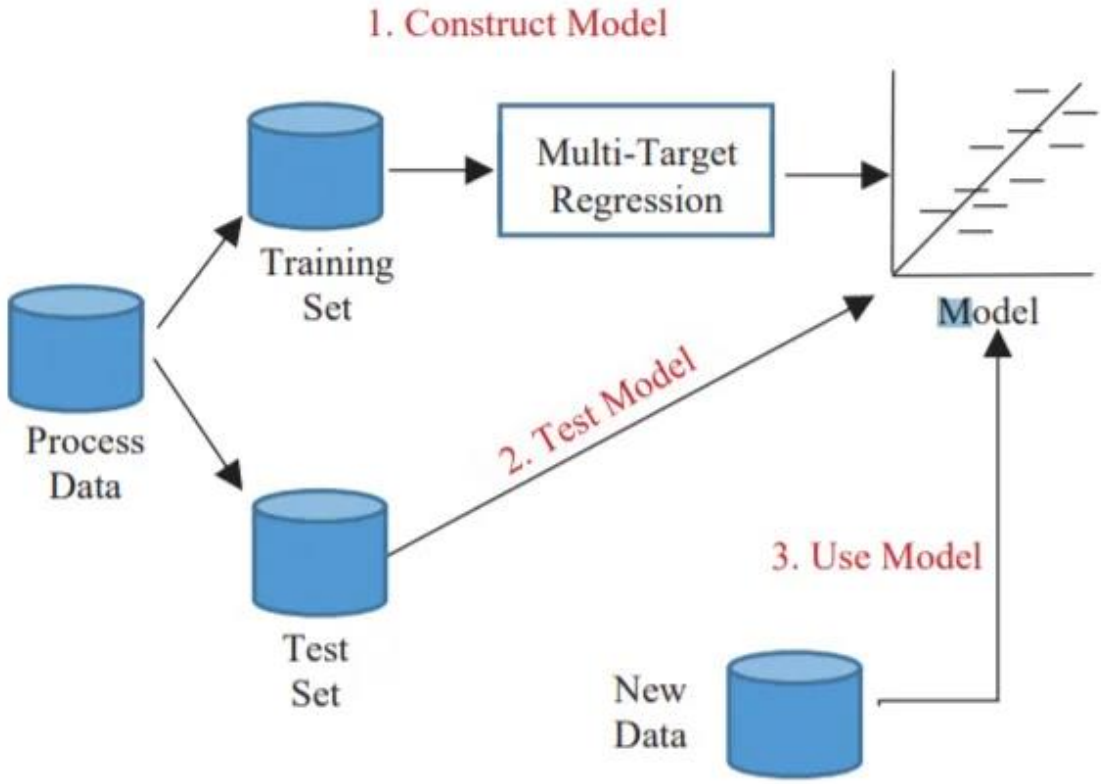
The experimental findings show that, in terms of coefficient of determination, the AdaBoost regressor regularly beats other methods. This is explained by the fact that AdaBoost uses an ensemble technique, which frequently improves accuracy by pooling the predictions of several predictors. To create a strong classifier from a collection of weak classifiers, AdaBoost iteratively modifies the weights of examples in the dataset with an emphasis on those that were incorrectly categorized. Random Forest, another ensemble learning method, also achieves a high score that is very near to the ideal number. As a result, Random Forest may be thought of as an alternative, particularly when dealing with many input variables, as it randomly picks feature subsets, shortening the runtime of the method. Although ensemble-based approaches (AdaBoost and Random Forest) produce superior accuracy, the findings also demonstrate adequate precision for the Decision Tree method. However, the Ridge technique performs noticeably poorer than decision tree-based approaches, making it unsuitable as a foundation learner for multi-output regression.

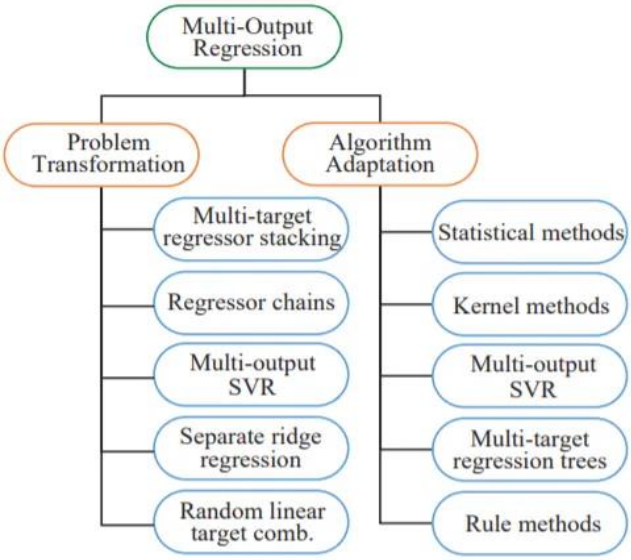
The goal of this work is to simultaneously address many targets in a multi-target regression issue to estimate the quality of a mining process.

4.3 High Level Diagram (if applicable)

none

4.4 Low Level Diagram (if applicable)





4.5 Interfaces (if applicable)

Update with Block Diagrams, Data flow, protocols, FLOW Charts, State Machines, Memory Buffer Management.

5 Performance Test

MODEL1	R2	MSE
ADABOOST	0.9745449086918032	0.03203487819099303
RANDOMFOREST	0.9404630125734039	0.0749937491528446
DECISION TREE	0.9036198036618404	0.12437333275470228
XGBOOST	0.7750269140015339	0.2641687929200151
RIDGE	0.1457070772811201	1.0722436207597437

The AdaBoost regressor fared better than other algorithms, according to the experimental results, in terms of the coefficient of determination. This is most likely because AdaBoost uses an ensemble technique, which combines the results of various weak models to get a single, more precise forecast. The algorithm repeatedly modifies the weights of the dataset's instances, paying particular attention to those who had inaccurate classifications in earlier iterations. Through this method, a powerful classifier can be created by combining several weak classifiers. The experiment's findings unequivocally show AdaBoost's advantage.

Random Forest, another ensemble learning method, received a high rating of 0.96, which is extremely near to the ideal score. As a result, Random Forest is an option as well, particularly when working with a high number of input variables. A subset of characteristics is chosen at random by Random Forest, which speeds up the algorithm's processing. Additionally, the outcomes demonstrate the Decision Tree method's respectable accuracy. AdaBoost and Random Forest, two ensemble-based approaches, perform better than this approach in terms of accuracy. On the other hand, because it performs noticeably poorer than decision tree-based techniques, the RIDGE method, as shown in the table, is unsuitable as a base learner for multi-output regression.

In this study, we attempted to forecast the quality of a mining process using a multi-target regression problem. Our goal was to create a reliable model

5.1 Test Plan/ Test Cases

The dataset (734543,24) is divided into a proportionate (80%,20%) train and test dataset.

Our major goal is to determine whether we can forecast silica without using iron concentrate.To carry out these action.

We can forecast the performance of the model by using R2 as a measure to create the model (with iron concentrate) using train data and predict the silica concentrate using test data utilizing train data to create a model (without iron concentrate), test data to predict the silica concentration, and R2 as an assessment metric.

Ridge retreat

When a data set exhibits multicollinearity (correlations between predictor variables) or when the number of predictor variables in the set is more than the number of observations, a model may be created using ridge regression. The attribute is the primary justification for picking ridge regression.

XGBREGRESSOR

Since the ridge regression exhibits lower performance metrics, the model's metrics are improved. Because of its increased performance and quicker performance, XGBoost is employed.

REGRESSOR FOR DECISION TREE

Using a graph that resembles a tree or model of decisions and their potential outcomes that include resource costs and utility, a decision tree is used. Only conditional control statements are used. Each internal node corresponds to a feature test. Every leaf node on the graph represents a class label or potential value of the item in regression, and every edge on the graph represents the result of the linked test. A 1D regression technique using a Decision Tree is used in the Scikit-Learn model.

Regressor from a Random Forest

On a variety of distinct subsamples, this technique essentially trains a number of categorizing decision trees. The averaging process helps to increase forecast accuracy and manage over-fitting. Replacement samples are chosen at random from training samples. Every new training batch has the same amount of data as the initial dataset. In other words, a chosen instance is likely to be selected repeatedly as a component of other subsets. The number of trees in the method and the maximum depth should first be defined as input parameters. The algorithm's performance and prediction ability may be impacted by the change in their values. As a result, the technique is evaluated with all feasible parameters within the range of the dataset size. The variables that produce the greatest results become the candidates for utilization. This technique works effectively without adding excessive computing overhead.

ADABOOST

Boosting is an ensemble technique that aims to create a powerful classifier from a collection of classifiers that are all relatively weak. The training data is used to create the model initially. After then, a second model is produced by fixing the previous model's mistakes. Models are built up gradually. A meta estimator called AdaBoost regressor begins by fitting a regressor on the initial dataset and then fits

several versions of the regressor on the same dataset. However, the record weights are changed in accordance with the most recent estimation inaccuracy.

MODEL COMPARISON (WITH AND WITHOUT CONCENTRATE IRON)

As we previously indicated, using the aforesaid models, we have examined whether we can estimate the percentage of silica with and without iron concentrate using R2 METRIC and MEAN-SQUARED-ERROR.

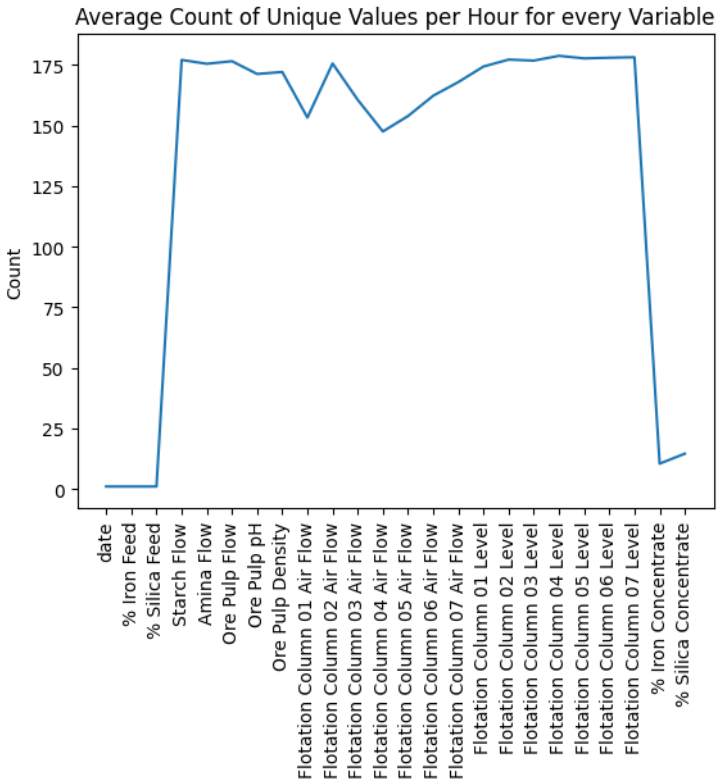
To assess the effectiveness of the techniques, the coefficient of determination (R2) statistic was utilized as the evaluation metric. It represents the potential for newly discovered data to fall within expected results.

5.2 Test Procedure

IMPLEMENTATION

EDA

During Eda Analysis of the dataset, check for %silica concentrate for one day and checking the variables which have hourly 20-sec frequency.



The diagram demonstrates how variables fluctuate over the course of an hour, although other variables, like iron and silica feed, barely move at all.

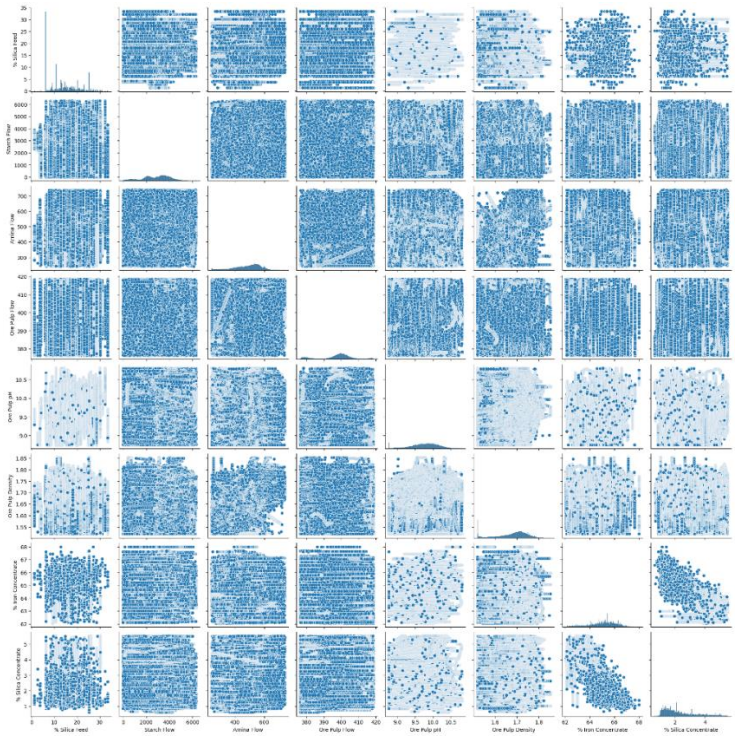
The scale is from 0 to 180 recordings since we sample once every 20 seconds, giving us 180 records altogether throughout the course of an hour.

The level and flotation air flow factors change every hour. The most crucial factors are therefore airflow and level. The yield rate of silica concentrate can be controlled by adjusting these factors.

Upon Pair plotting for gathering information for correlated data:-

There are no significant patterns observed, except for the expected correlation between iron and silica concentrate, as well as iron and silica feed. Additionally, we conducted an analysis to determine which minute of the hour showed the highest correlation with the percentage of silica concentrate for each variable. The hypothesis was that these correlations would peak around the time when the measurements were typically taken.

Furthermore, we noticed a negative correlation between the iron and silica concentrate. It was observed that when the iron ore contains a large amount of iron concentrate, the impurities (silica concentrate) are minimized. This finding suggests that the iron ore used in our process is of high quality.



End Conclusions from EDA Analysis:-

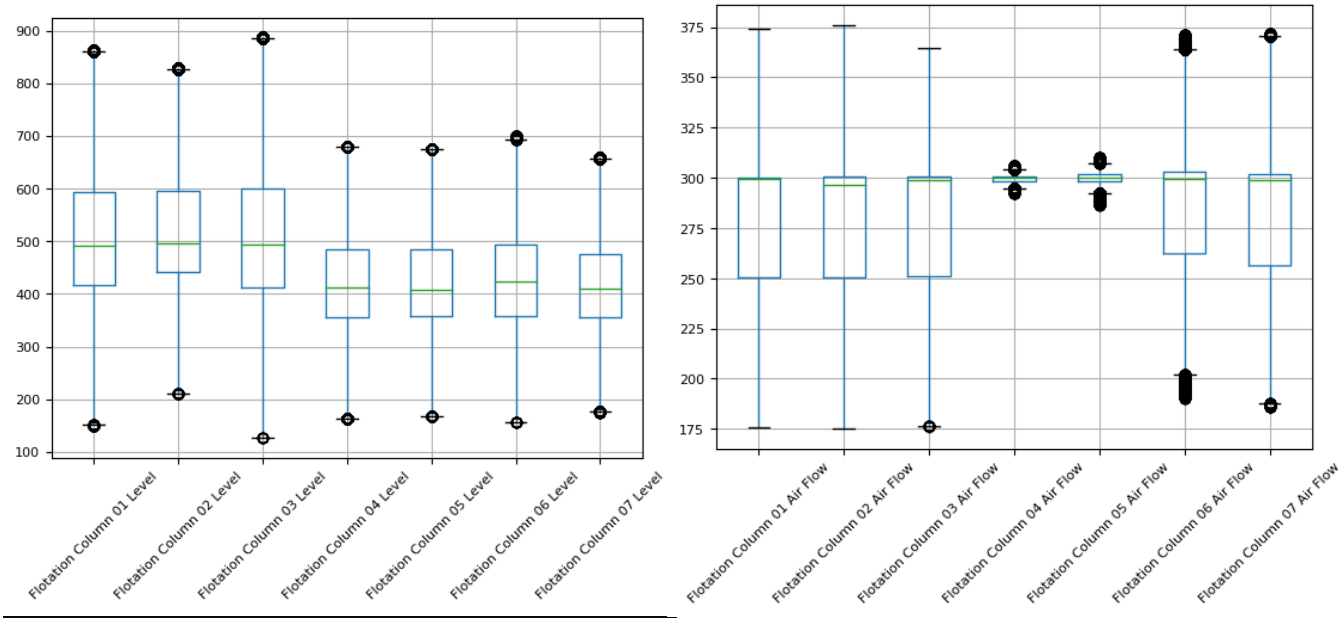
. In high level statistics,1. The shape of dataset is (737453,24) which means the dataset has total 737453 data samples generated every 20 secs from march 2017 to September 2017 and there are 24 features to determine the percentage of silica concentrate .

2. With the help of these percentiles we can draw to conclusion for increase in the feed rates we can see the increase in % silica concentrate and iron concentrate

3. From the trend of % silica concentrate graph we get the most important variables are air-flow and level. Controlling these variables can be used to control the yield rate of silica concentrate.

4. We can conclude from iron feed graph that we more the iron feed greater is the silica concentrate.

5. We are making very important conclusion from iron concentrate vs silica concentrate plot that both the variables are negatively correlated.



2. Data Preprocessing(Missing value treatment)

We must calculate the necessary number of records using the formula below in order to identify the missing values:

1 Hour = 3600 Samples at 20 Seconds in a Session

We can get the required number of records by assuming that one record ends every 20 seconds:

3600/20 = 180

In other words, we get 180 recordings in an hour. If we discover that there are 180 records in all, we can say there are no missing data. If the number of records is different, we must fill in the blanks with the appropriate values.

```
counts = df.groupby('date').count()
counts
```

Python

date	% Iron Feed	% Silica Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 02 Air Flow	Flotation Column 03 Air Flow	...	Flotation Column 07 Air Flow	Flotation Column 01 Level	Flotation Column 02 Level	Flotation Column 03 Level	Flotation Column 04 Level	Flotation Column 05 Level	Flotation Column 06 Level	Flotation Column 07 Level	% Iron Concentrate	Co
2017-03-10 01:00:00	174	174	174	174	174	174	174	174	174	174	...	174	174	174	174	174	174	174	174	174	
2017-03-10 02:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
2017-03-10 03:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
2017-03-10 04:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
2017-03-10 05:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2017-09-09 19:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
2017-09-09 20:00:00	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	
2017-09-09	180	180	180	180	180	180	180	180	180	180	...	180	180	180	180	180	180	180	180	180	

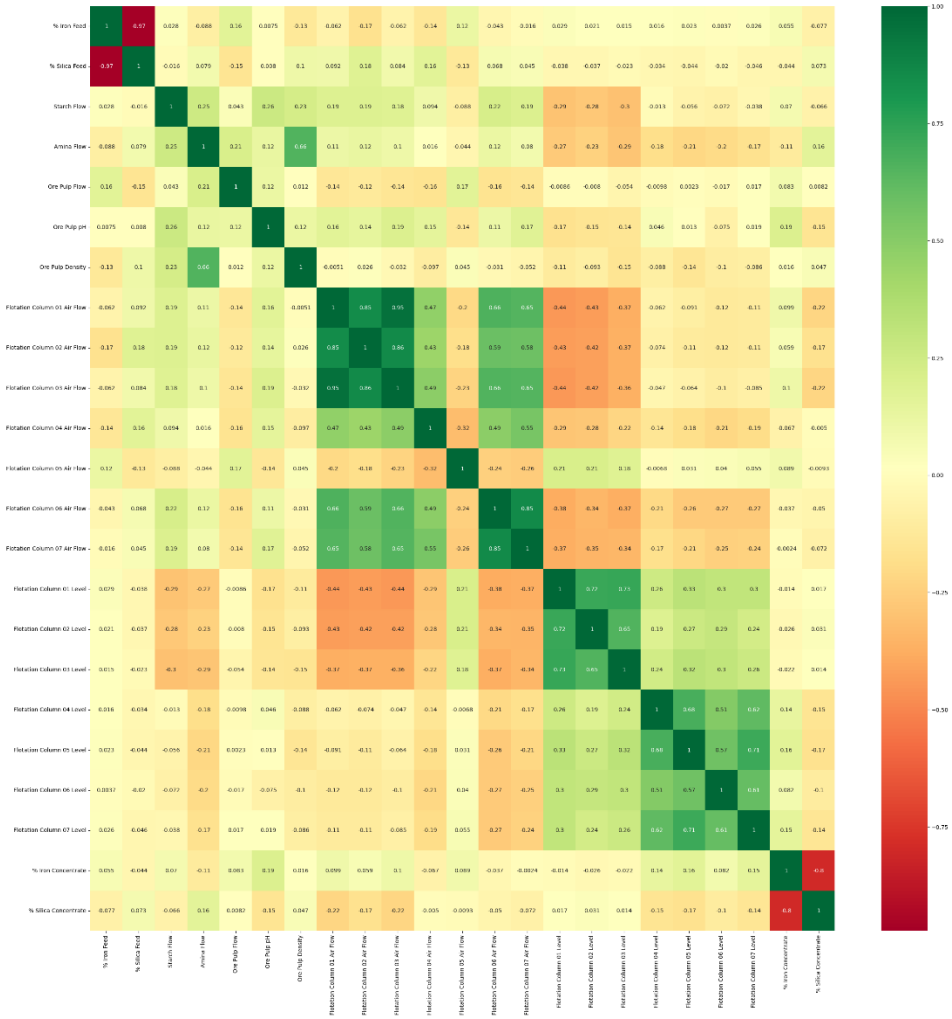
NULL VALUE DETECTION

There is no null values.

CORRELATION

We discovered numerically that there is a negative correlation between the concentrations of iron and silica using specific Python techniques.

They have a negative correlation, as seen in the heatmap. In other words, silica content in iron ore reduces as iron content increases.



Artificial neural networks with representation learning are used in deep learning, a branch of machine learning. It includes a variety of topologies, including convolutional neural networks, recurrent neural networks, deep belief networks, and neural networks with recurrent connections. Deep learning has been successfully used in a variety of domains, including computer vision, speech recognition, natural language processing, audio recognition, and more, frequently producing outcomes that are on par with or better than those produced by experts in those fields.

We use deep learning models in our case study to forecast the desired variables. Our methodology is outlined in the steps below:

- Splitting the Dataset: The dataset is split into two sets, D1 and D2.

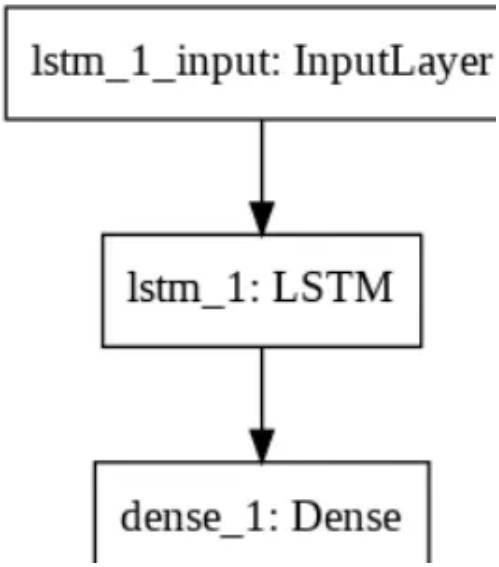
- Sampling with Replacement: With sampling with replacement, N sets of samples are produced, each having the same size as the original dataset.
- Model Construction: Using the datapoints from D1, N models are built.
- Each model is used to forecast the datapoints from D2.
- The predictions from the previous stage are fed into the deep learning model to forecast the final output.

LSTM

For our deep learning model, we use a straightforward design. The Long Short-Term Memory (LSTM), a kind of recurrent neural network (RNN), is a famous architecture. Both individual data points and data sequences can be processed by LSTM, which contains feedback connections. It works especially well for jobs involving linked, unsegmented handwriting recognition, speech recognition, and network traffic anomaly detection (IDS). A cell, an input gate, an output gate, and a forget gate are all components of the LSTM unit, which enables it to remember values over time and control the flow of information.

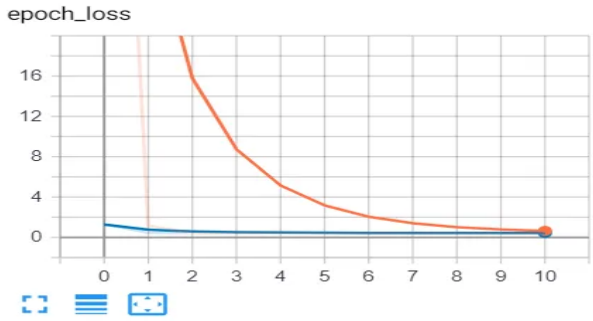
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_2 (LSTM)                (None, 32)                4480
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 66
=================================================================
Total params: 4,546
Trainable params: 4,546
Non-trainable params: 0
_________________________________________________________________
```

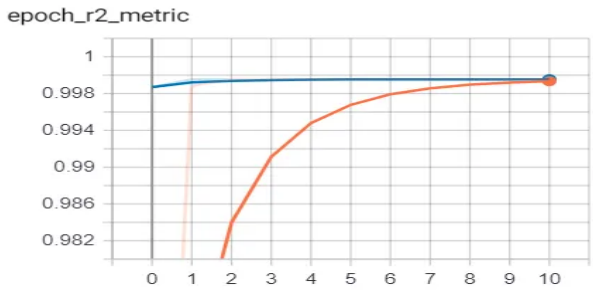
Based on time series data, LSTM networks are excellent at categorizing, analyzing, and making predictions.



The vanishing gradient issue that traditional RNNs frequently experience is avoided by them since they can accommodate time delays of various lengths between significant events. Furthermore, LSTM models demonstrate relative gap length insensitivity, giving them an advantage over other sequence learning techniques like hidden Markov models.

We use loss functions and the R2 metric, which assess the model's effectiveness and the percentage of variation explained by the predictions, for further examination of the model.





5.3 Performance Outcome

For this project, we attempted to forecast the quality of a mining process using a multi-target regression problem. Our objective was to create a reliable model that could calculate the proportions of silica and iron concentrations in the ore at the same time. To manage many target variables, we put various strategies into practice and compared them. We specifically looked at how well a multi-target regression method performed when the target features showed substantial correlation.

We found that this method can still produce useful data even when an associated attribute isn't passed as an input parameter to the algorithm. Instead, by adding it to the current target feature, this characteristic can be used as an output variable. Importantly, this change had no adverse effects on the regression performance.

In the end, the experimental findings demonstrated the AdaBoost regressor's higher performance.

6 My learnings

During my internship on the project "Quality Prediction in Mining Process," I have gained valuable insights and learnings that have significantly contributed to my growth in the field of data science and machine learning. Here are my key learnings from this internship:

- Domain Knowledge: Working on this project has deepened my understanding of the mining industry and the intricacies of quality prediction in mining processes. I have acquired knowledge about the various parameters and factors that impact the quality of mining products, which has enhanced my ability to approach data analysis and modeling from a domain-specific perspective.
- Data Preprocessing: The internship has provided me with hands-on experience in collecting, cleaning, and preprocessing real-world mining process data. I have learned to address challenges such as missing values, outliers, and data inconsistencies. The importance of data preprocessing in ensuring accurate and reliable models has become evident through this practical experience.
- Feature Engineering: Through feature engineering, I have learned how to extract meaningful information from raw data and transform it into informative features that improve the predictive power of models. Exploring different feature engineering techniques and evaluating their impact on model performance has expanded my knowledge of feature selection, transformation, and creation.
- Model Development and Optimization: The internship has allowed me to gain expertise in selecting appropriate algorithms and developing machine learning models for quality prediction. I have learned to implement and evaluate various models, optimizing their hyperparameters and evaluating their performance using evaluation metrics such as R^2 score. This experience has sharpened my skills in model development, tuning, and selection.
- Collaboration and Communication: Working on a self-guided internship has taught me the importance of self-discipline, time management, and effective communication. I have learned to collaborate with project stakeholders, seek guidance when needed, and present my findings and progress in a clear and concise manner. These skills are crucial for successful project execution and building professional relationships.
- Problem-solving and Adaptability: Throughout the internship, I encountered challenges such as limited data availability and complexity in feature engineering. These hurdles have honed my problem-solving skills and taught me to adapt my approaches based on the specific project requirements and constraints. I have learned to think critically, experiment, and iterate to find effective solutions.

Overall, this internship has been a transformative learning experience, providing me with practical skills and knowledge in data science and machine learning. I am confident that the learnings from this project, including domain expertise, data preprocessing, feature engineering, model development, collaboration, and problem-solving, will significantly contribute to my future endeavors in the field of data science and positively impact my professional growth.

7 Future work scope

We used LSTM as a DL model and predicted the outcomes. We can also predict outcomes using CNN+LSTM.
For other manufacturing processes where variables are highly correlated, we can attempt to apply MTR.