Industrial Internship Report on

”Quality Prediction In Mining Process”

Prepared by

Parvath Darshini S

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 data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, so 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 to reduce impurity and also help the environment. The key analysis done in this project are to predict % Silica Concentrate every minute, how many hours ahead can we predict % Silica in Concentrate and to predict % Silica in Concentrate without using % Iron Concentrate column. 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: -

This report provides a comprehensive summary of a six-week internship focused on quality prediction in mining. The document will outline the need for relevant internships in career development, provide a brief overview of the project and problem statement, discuss the opportunities provided by the USC and UCT, and describe how the program was planned and executed.

- **Summary of the Whole Six Weeks' Work:**

The six-week internship was structured to provide a practical and in-depth understanding of quality prediction in mining. The primary objective was to develop and implement a model that could predict the quality of mined materials, ensuring efficient resource management and reducing operational costs.

- Week 1: Orientation and Introduction
- Week 2: Data Collection and Preprocessing
- Week 3: Model Development
- Week 4: Model Evaluation and Optimization
- Week 5: Implementation and Testing
- Week 6: Maintenance and Reporting

- **The Need for Relevant Internships in Career Development:**

Relevant internships play a crucial role in career development by providing hands-on experience, industry exposure, and practical skills that complement academic learning. They offer students the opportunity to apply theoretical knowledge in real-world settings, thereby bridging the gap between academia and industry. Internships also facilitate networking, mentorship, and professional growth, making them an indispensable component of career development.

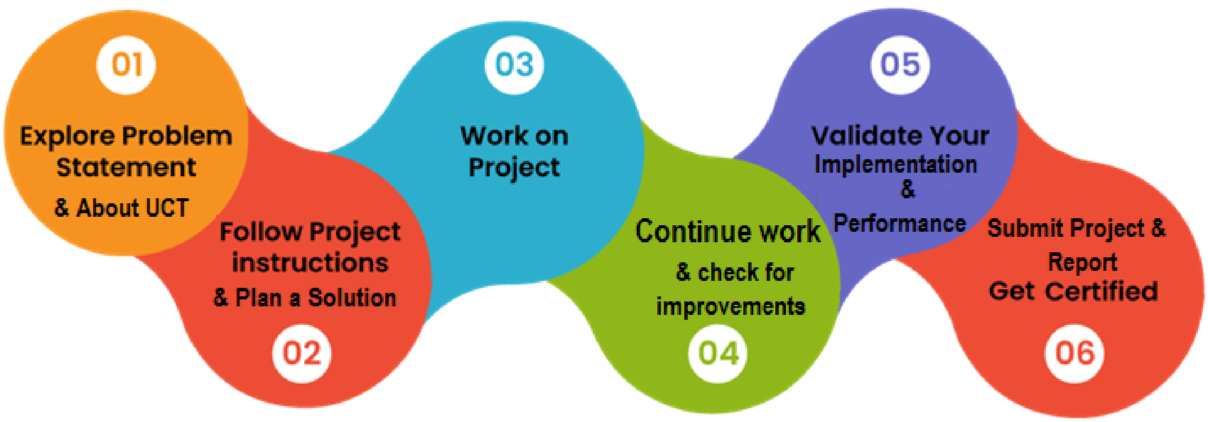
- **Brief About Your Project/Problem Statement:**

The project aimed to develop a predictive model for determining the quality of mined materials. Accurate quality prediction is vital for efficient resource allocation, cost reduction, and ensuring the sustainability of mining operations. The problem statement was to leverage historical data and machine learning techniques to predict the quality of materials extracted from mining sites accurately.

- **Opportunity Given by USC/UCT:**

The internship was a collaborative effort between the Upskills Campus (USC) and the Uniconverge Technologies (UCT). Both platforms provided valuable resources, including access to data, expert guidance, and state-of-the-art computational facilities. The interdisciplinary nature of the program allowed for a rich exchange of ideas and expertise, significantly enhancing the learning experience

- **How the Program Was Planned:**



1. **Problem Exploration & Understanding**:
   - Begin by thoroughly understanding the problem statement related to quality prediction in mining.
   - Identify the specific aspects of quality that need to be predicted (e.g., product quality, process quality, safety, etc.).
2. **Project Guidelines & Solution Planning**:
   - Follow any provided project instructions or guidelines.
   - Strategize a solution approach based on the problem requirements.
3. **Project Execution:**
   - Actively work on implementing the solution.
   - Consider data collection, feature engineering, and model development.
4. **Continuous Improvement & Refinement:**
   - Regularly review the project's progress.
   - Look for opportunities to enhance the solution (e.g., fine-tuning model parameters, incorporating additional features).
5. **Validation of Implementation**:
   - Evaluate the performance of the implemented solution.

- Use appropriate evaluation metrics to assess the quality prediction accuracy.

6. **Certification Process:**
    - Prepare a comprehensive report summarizing the project.
    - Submit the project for certification, demonstrating the successful prediction of quality in the mining context.

- **Overall Learning and Experiences: -**

My internship at UCT and USC Mining was both challenging and rewarding. The blend of technical skills and industry insights enriched my understanding of mining operations. I'm grateful for the hands-on experience and the chance to contribute to improving quality standards in mining.

This internship has been an invaluable learning experience. I particularly appreciate Nitin Sir and the other interns who helped to solve my doubts.

Thanks to this experience, I feel more confident in my abilities in Machine Learning, Probability and statistics and have gained a deeper understanding of Mining Industry. I am excited to take these newfound skills and knowledge with me for upcoming internships and career growth.

Thank you again for everything.

1 Introduction

1.1 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)

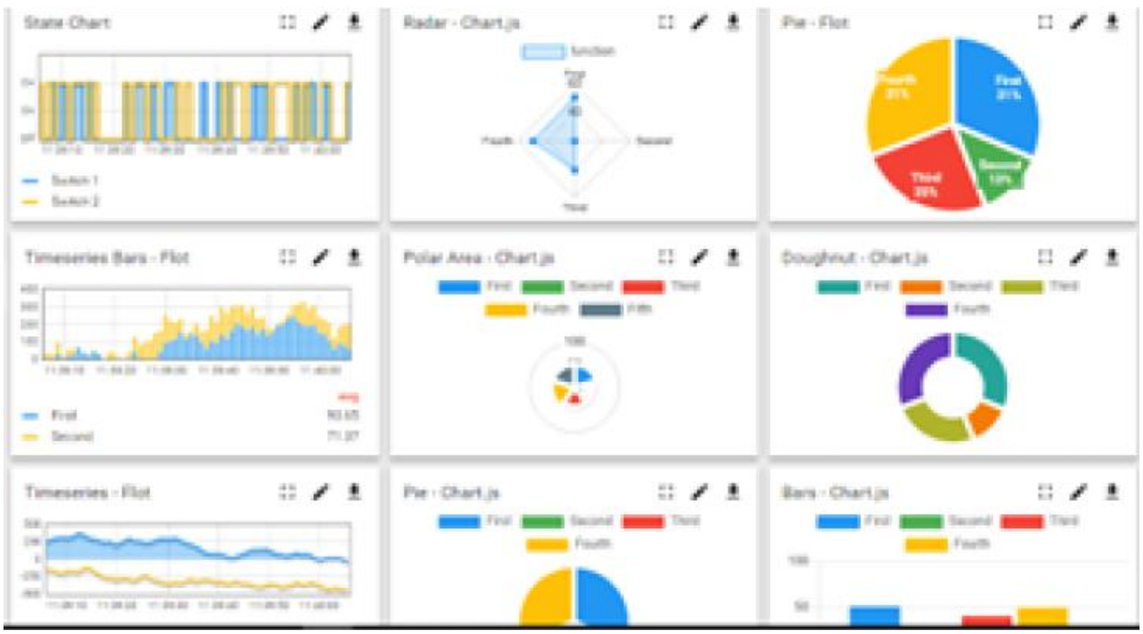
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

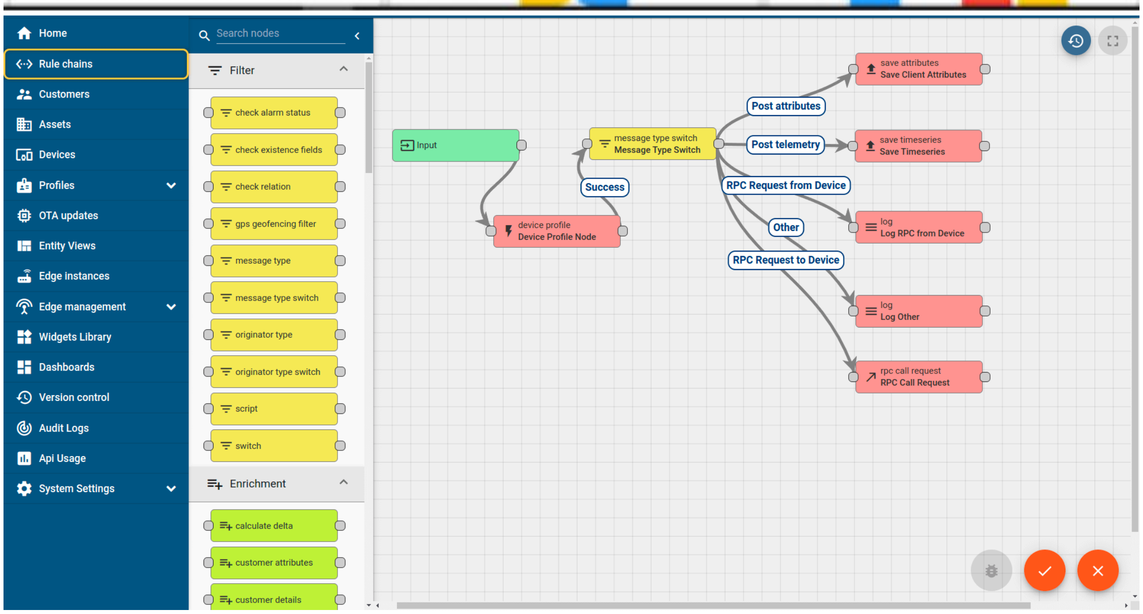
- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA

- 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.

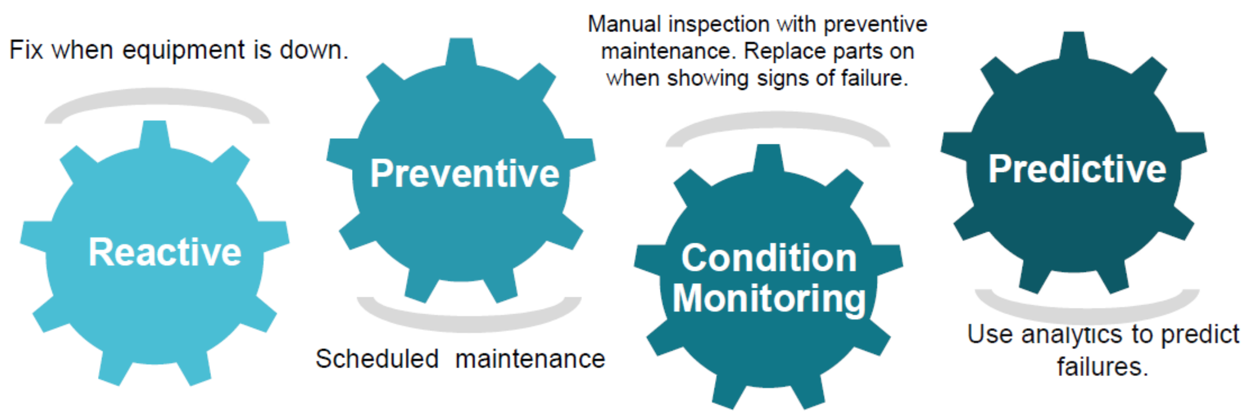
Its unique SaaS model helps users to save time, cost and money.

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

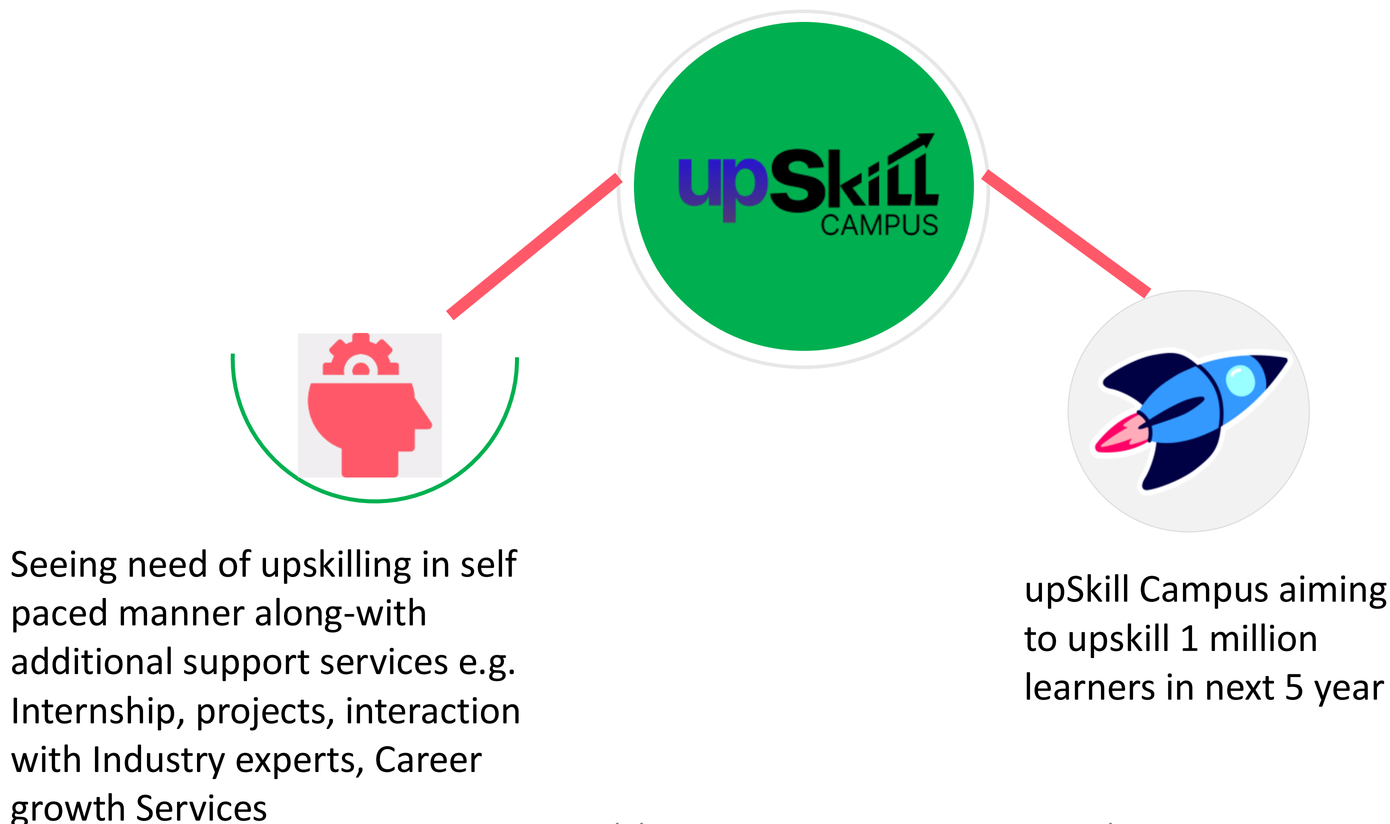
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



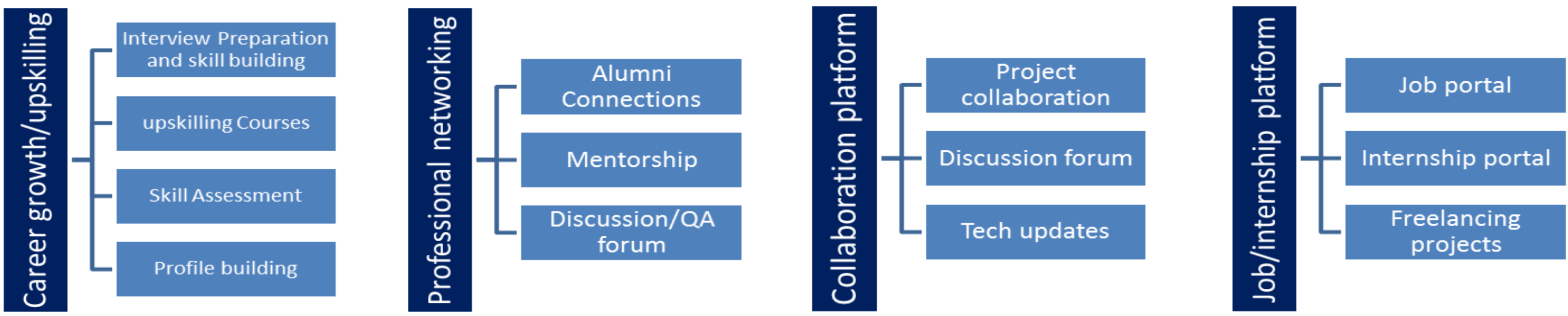
1.2 About upskill Campus (USC)

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.



https://www.upskillcampus.com/



1.3 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.

1.4 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.

1.5 Reference

[1] Quality Prediction in a Mining Process | by TechLabs Aachen | Medium

[2] Machine Learning based quality prediction for milling processes using internal machine tool data - ScienceDirect

[3] Online Course: Mining Quality Prediction Using Machine & Deep Learning from Coursera Project Network | Class Central

1.6 Glossary

Terms	Acronym
Quality Prediction	Forecasting the quality characteristics of mined materials or products based on various parameters and historical data.
Mining Operations	Activities involved in extracting minerals from the earth, including exploration, excavation, processing, and transportation.
%Silica in Concentrate:	The percentage of silica content in the final concentrate product after mineral processing.
Predictive Modeling	Using statistical and machine learning techniques to predict outcomes based on historical data, applied in mining to forecast ore quality parameters.
Data Preprocessing	Cleaning, transforming, and normalizing raw data before applying predictive models, ensuring data quality and compatibility with modeling techniques.

2 Problem Statement

Project Statement: Quality Prediction in Mining

1. **Introduction:**

In the mining industry, the ability to accurately predict the quality of mined materials is crucial for operational efficiency, cost management, and ensuring product standards. Quality prediction in mining involves the use of various techniques and tools to forecast the characteristics and value of materials extracted from the earth. This project aims to develop a robust model to predict the quality of mined resources, thereby optimizing the extraction process and improving decision-making.

2. **Objectives**

The primary objectives of the project are:

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 (empowering!). 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.

I have tried to answer the following questions through my model:

- Is it possible to predict % Silica Concentrate every minute?
- How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.
- Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?

3. **Methodology**

The methodology for this project involves several key steps:

- **Data Collection:** Collect data from various sources. In my model, the dataset is provided by the company.

- **Data Cleaning and Preprocessing:** Clean the data to handle missing values, outliers, and noise. Normalize and transform the data as needed to prepare it for analysis. I have done cleaning dataset through Excel and Pandas.
- **Exploratory Data Analysis (EDA):** Perform EDA to understand the underlying patterns and relationships within the data. Visualize key features and their correlations with material quality. Heatmap is used for EDA in my model.
- **Feature Engineering:** Select and engineer features that are predictive of quality. This may include geological characteristics, environmental conditions, and operational parameters. I have used scatter plot irrelevant features are identified and dropped.
- **Model Development:** Use machine learning algorithms to develop predictive models. I have used five ML algorithms i.e. Linear Regression Model, XGBoost Model, Decision Tree, Lasso Regression Model and Ridge Model.
- **Model Training and Validation:** Train the models on historical data and validate their performance using cross-validation techniques. I have used 70% of dataset for training the model.
- **Model Evaluation:** Evaluate the models using metrics. I have used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), Mean Squared Error (MSE) to determine their accuracy and reliability.
- **Testing:** Testing the models to ensure the validation of the data and building high quality models. I have used 30% of dataset into testing phase.

4. Expected Outcomes:

By the end of this project, we expect to achieve the following outcomes:

- Build a model to predict % Silica Concentrate every minute.
- Build a model that will show how many steps (hours) ahead can we predict % Silica in Concentrate so that This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.
- Model to predict % Silica in Concentrate without using % Iron Concentrate column.

3 Existing and Proposed solution: -

 # Summary of Existing Solutions in Quality Prediction in Mining and its Limitations: -

1. Traditional Statistical Methods: Traditional statistical methods such as linear regression and time series analysis have been used to predict the quality of mined materials based on historical data.

Limitations:

- Simplicity: While easy to implement, these methods often oversimplify complex relationships between variables.
- Limited Scope: They typically do not account for non-linear relationships and interactions between multiple factors.
- Data Dependency**:** High dependency on the quality and quantity of historical data, making them less effective when data is scarce or incomplete.

2. Geostatistical Methods: Geostatistical techniques like kriging and cokriging are used to model and predict spatial variability in ore quality based on sample data from different locations.

Limitations:

- Spatial Focus**:** Primarily focused on spatial relationships, sometimes overlooking temporal dynamics and operational factors.
- Computational Complexity**:** Can be computationally intensive, especially for large datasets with high spatial resolution.
- Accuracy**:** Prediction accuracy heavily depends on the spatial density and distribution of sample points.

3. Machine Learning Models: Machine learning models, including decision trees, random forests, support vector machines (SVM), and neural networks, have been increasingly used to predict mining quality by learning complex patterns in data.

Limitations:

- Data Quality**:** These models require large amounts of high-quality, labeled data for training, which may not always be available.

- Interpretability: Many machine learning models, especially deep learning models, are often considered black boxes, making it difficult to interpret and trust their predictions.
- Overfitting: There is a risk of overfitting, where the model performs well on training data but poorly on unseen data.
- Computational Resources**:** High computational resources and expertise are required to develop and maintain these models.

4. Expert Systems: Expert systems utilize rule-based approaches where domain knowledge from mining experts is encoded into the system to predict ore quality.

Limitations:

- Static Rules: The rule-based nature means the system may not adapt well to new or changing conditions.
- Scalability: Difficult to scale as the complexity of mining operations increases.
- Expert Dependency**:** Heavy reliance on the availability and accuracy of expert knowledge, which can be subjective and limited.

5. Hybrid Approaches: Combining multiple methods, such as integrating geostatistical methods with machine learning, to leverage the strengths of each approach for better prediction accuracy.

Limitations:

- Complexity: Increased complexity in model development and implementation.
- Integration Challenges: Challenges in effectively integrating different models and ensuring they work together seamlessly.
- Maintenance: Higher maintenance requirements due to the complexity and number of components involved.

 # Proposed Solution for Quality Prediction in Mining: -

- **Key Components of the Proposed Solution:**

1. Data Integration and Preprocessing
2. Advanced Machine Learning Models
3. Continuous Learning and Adaptation

4. Scalable Infrastructure
5. User-friendly Interface
6. Collaboration with Domain Experts

- The proposed solution for quality prediction in mining combines the strengths of advanced machine learning models, geostatistical techniques, and domain expertise to overcome the limitations of existing methods. By providing accurate, scalable, and interpretable predictions, this solution aims to optimize mining operations, reduce costs, and improve resource management, ultimately leading to more efficient and profitable mining activities.

 # Value Addition of the Proposed Solution:

1. Improved Data Management

- Data Integration: Combining data from different sources (e.g., geological surveys, production records, sensor readings) to have a complete and comprehensive dataset.
- Data Cleaning: Ensuring that the data is accurate and free from errors, such as missing values or incorrect entries.

2. User-friendly Tools

- Visualization Tools: Providing easy-to-understand charts and graphs that show the quality predictions and key trends, helping users make informed decisions quickly.
- Simple Dashboard: Creating an interface where users can easily see the predictions, historical data, and important metrics all in one place.

3. Better Prediction Methods

- Multiple Methods: Using different simple prediction methods and comparing their results to choose the best one, like using both linear regression and basic decision trees.
- Combining Techniques: Using basic techniques like combining average predictions from multiple models to get a more reliable result.

4. Real-time Updates

- Real-time Data: Using current data to make predictions immediately, so users can act on the latest information.
- Alerts and Notifications: Setting up alerts to notify users of significant changes or predictions that require attention.

5. Collaboration with Experts

- Expert Input: Working with mining experts to ensure the predictions are realistic and practical.
- Field Validation: Regularly checking predictions against actual field data to confirm accuracy and improve the model.

6. Continuous Improvement

- Regular Updates: Periodically updating the prediction methods with new data to keep them accurate and relevant.
- Feedback Loop: Allowing users to provide feedback on the predictions, which can be used to refine and improve the methods.
- The proposed solution adds value by improving data management, providing user-friendly tools, employing better prediction methods, offering real-time updates, collaborating with experts, and continuously improving the system. These straightforward improvements lead to more accurate predictions, better decision-making, and more efficient mining operations.

3.1 Code submission (Github link):

" Quality-Prediction-Of-Mining/mining project.py at main · Parvathdarshini/Quality-Prediction-Of-Mining (github.com) "

3.2 Report submission (Github link):

" "

4 Proposed Design/ Model

 # Design Flow of a Solution: -

1. Problem Definition and Understanding:

- Identify and define the problem statement clearly. Understand the business or domain context where the solution will be applied. Then import the required libraries, as I have used python I have imported Pandas, Matplotlib, Seaborn, Numpy and Sklearn.
- Clearly define the objectives and success criteria for the project. Establish metrics for evaluating model performance.

2. Data Collection and Preparation:

- Gather relevant data sources needed to address the problem statement. Like for this project the dataset is provided by the company itself.
- Cleanse the data by handling missing values, outliers, and inconsistencies. Perform data preprocessing steps such as normalization, feature scaling, and encoding categorical variables. There were many outliers and inconsistencies in the data and I have cleaned it by using Excel and pandas and formatting the data. Then changed the datatype from object to float.
- Prepare a clean, structured dataset that is ready for analysis and modelling.

3. Feature Engineering and Selection:

- Generate new features from existing data that may improve model performance. Perform Exploratory Data Analysis (EDA) to find the relevant features and keep it only i.e., drop the irrelevant features.
- Select relevant features that have the most predictive power using techniques like statistical tests, feature importance rankings, or domain knowledge. I have used Heatmap and scatter plots to find the correlations between the features and find the irrelevant features.
- I have dropped the following features which are irrelevant to our model : '%IronFeed','StarchFlow','OrePulppH','FlotationColumn02AirFlow','FlotationColumn03AirFlow', 'FlotationColumn04AirFlow','FlotationColumn05AirFlow','FlotationColumn06AirFlow', 'FlotationColumn07AirFlow','FlotationColumn01Level','FlotationColumn02Level', 'FlotationColumn03Level','FlotationColumn05Level','FlotationColumn06Level', 'FlotationColumn07Level.

- Finalize the set of features to be used for model training and validation.
- Then scale the data and split the data for training and testing. I have splitted the dataset into 70% for training and 30% for testing.

4. Model Selection and Training:

- Choose appropriate algorithms and models based on the problem type (e.g., classification, regression) and data characteristics. Here I have used five models to train the dataset, they are Linear Regression, Decision Tree, XG Boost, Lasso and Ridge Model.
- Train initial models using training data and evaluate their performance using validation techniques such as cross-validation.
- Select the best-performing model(s) based on evaluation metrics and refine them further if necessary.

5. Model Evaluation and Testing and Tuning:

- Assess model performance on unseen data (test set) to validate its generalization capability.
- Fine-tune hyperparameters to optimize model performance using techniques like grid search, random search, or Bayesian optimization.
- Validate the final model's performance using appropriate evaluation metrics and ensure it meets business objectives. I have used 4 metrics Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 Score ((R^2), Root Mean Squared Error (RMSE).
- I have used random search and found the best model i.e. in my model its XGB Regressor.

6. Documentation and Reporting:

- Document the entire design process, including data sources, methodologies, and implementation details.
- Provide comprehensive documentation for future reference and knowledge transfer. Ensure transparency in the decision-making process and the implications of the model's outcomes.

7. Maintenance and Iteration:

- Establish procedures for maintaining and updating the deployed model to ensure continued relevance and accuracy.

- Monitor model performance in real-world scenarios and collect feedback for continuous improvement.
- Iterate on the solution based on feedback and evolving business requirements, adapting to new challenges and opportunities.

4.1 High Level Diagram (if applicable)

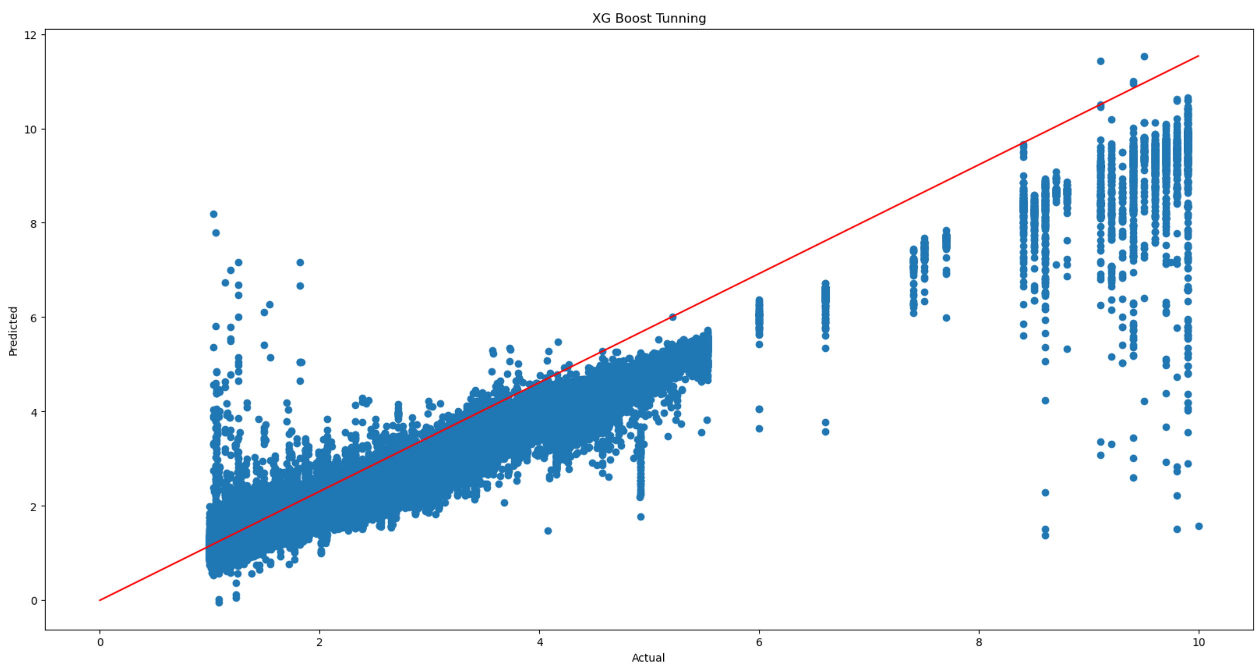
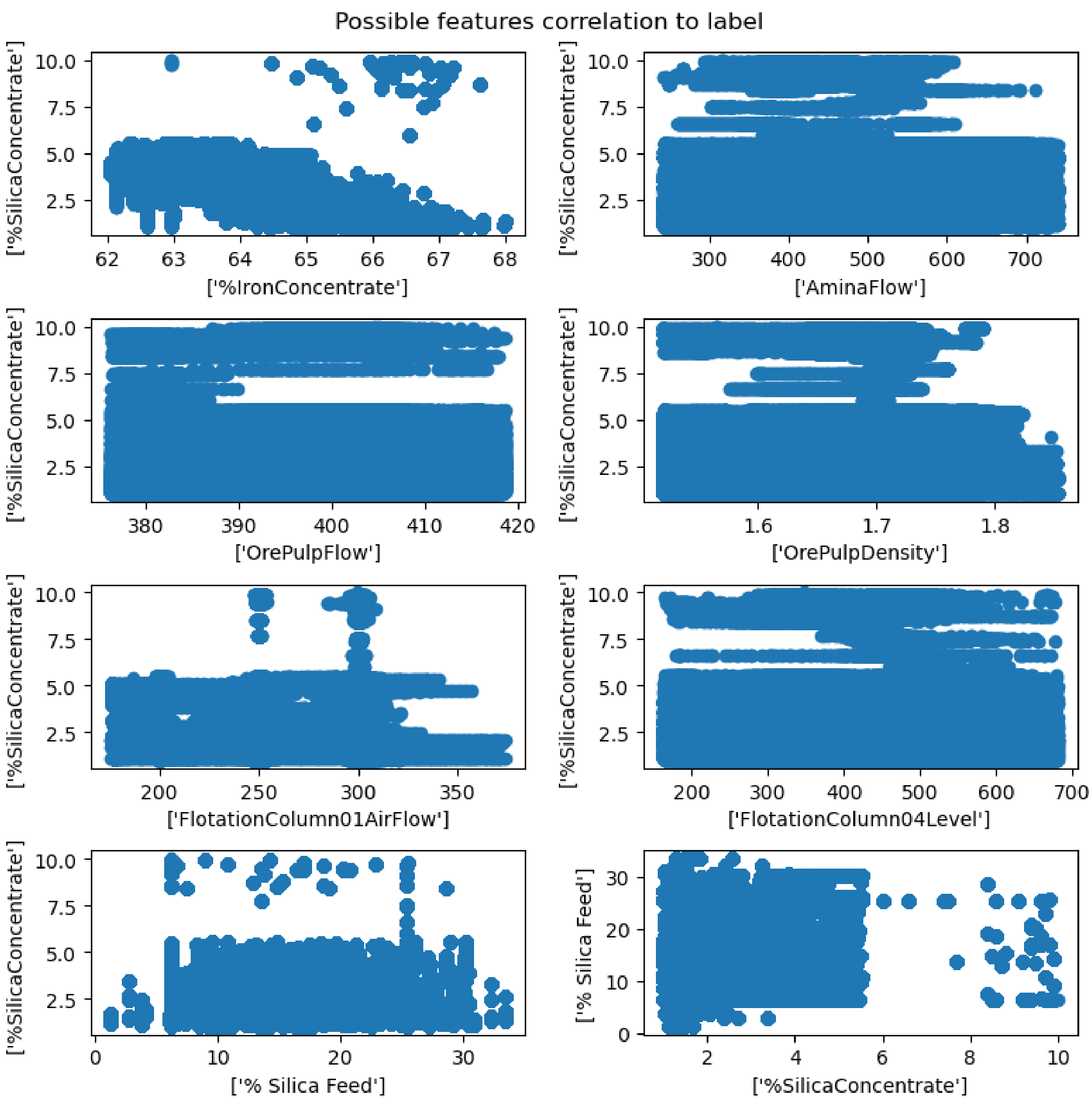
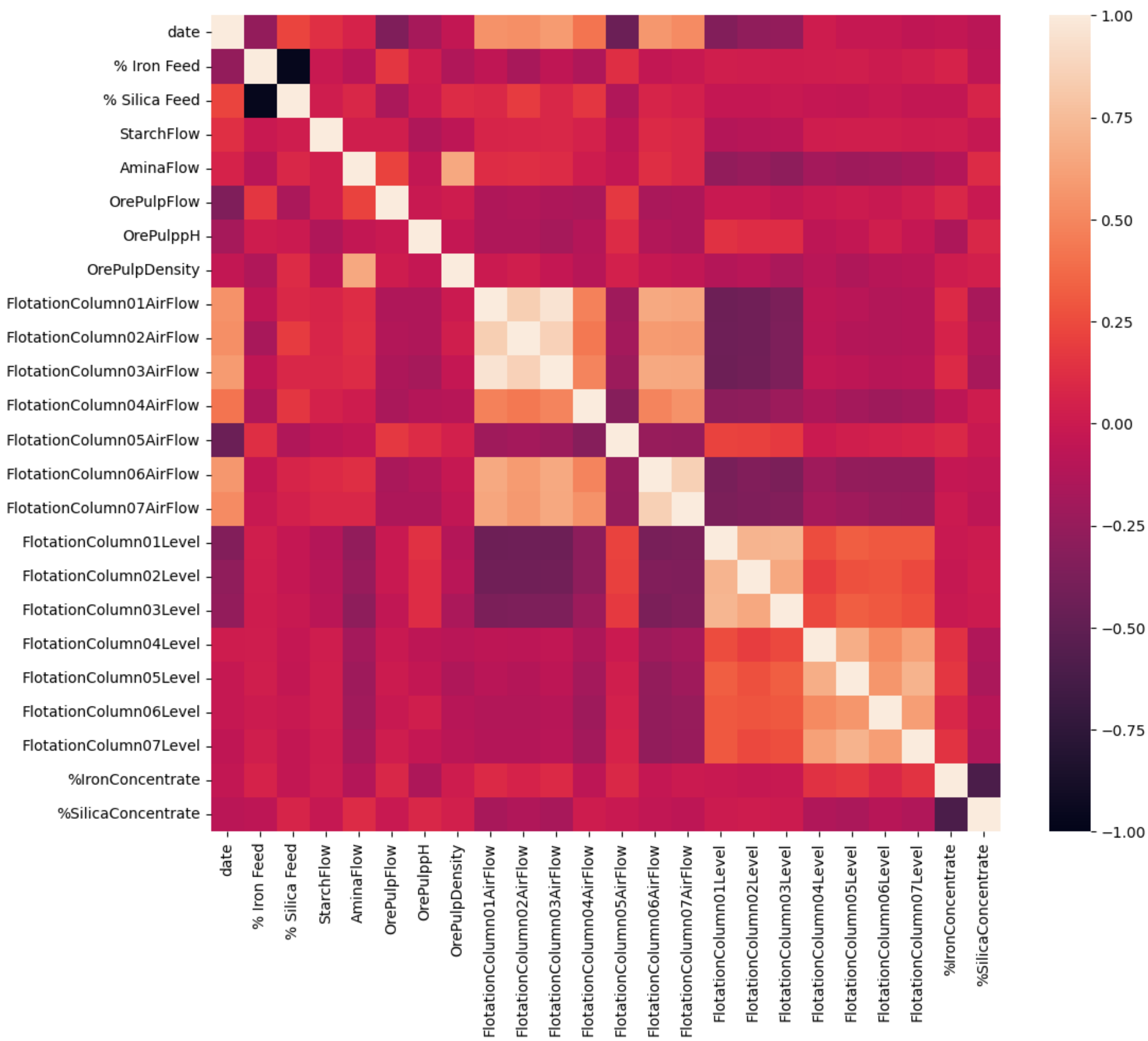


Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM

Here I have used XG Boost tuning Predicted value and Actual value. It is the final performance of my model.

4.2 Low Level Diagram (if applicable)





The above two diagrams are used to EDA and feature engineering so that the correlation between the features.

5 Performance Test

 # Identifying Constraints: -

1. **Time Constraints:**
    - Identification: Recognizing limited project duration and deadlines for deliverables.
    - Impact: Influenced project scope and depth of analysis.
    - Addressed in Design: Prioritized essential tasks and focused on achievable goals within the given timeframe. Leveraged agile methodologies to iterate and adjust as needed.
2. **Data Availability and Quality:**
    - Identification: Assessing data completeness, accuracy, and reliability.
    - Impact: Directly affects the accuracy and reliability of predictive models.
    - Addressed in Design**:** Conducted thorough data preprocessing, including cleaning, normalization, and outlier detection. Implemented robust validation techniques to handle potential biases and missing data.
3. **Resource Limitations:**
    - Identification: Considering constraints on computational resources, budget, and personnel.
    - Impact: Limits scalability and complexity of models and analyses.
    - Addressed in Design: Optimized algorithms and workflows for efficiency. Utilized cloud computing or distributed processing to manage large datasets. Maximized existing resources through effective use of open-source tools and platforms.
4. **Regulatory and Environmental Factors:**
    - Identification: Understanding legal requirements, environmental regulations, and safety standards.
    - Impact: Influences operational practices and data collection methods.

- Addressed in Design**:** Incorporated compliance checks into data handling and model development processes. Considered environmental impact assessments and sustainability metrics in predictions and recommendations.

 # Addressing Constraints in Design: -

1. **Methodological Adaptation:**

   - Approach: Chose methodologies and models that fit within resource and time constraints, such as simpler models initially with the potential for scalability.

   - Implementation: Used scalable data processing frameworks and optimized model training processes to manage resource limitations effectively.

2. **Data Handling Strategies:**

   - Data Preprocessing: Implemented robust preprocessing pipelines to handle data quality issues upfront, ensuring models are trained on reliable data.

   - Feature Selection: Employed feature selection techniques to focus on relevant data attributes, reducing computational overhead and improving model performance.

3. **Validation and Testing:**

   - Validation Techniques: Employed cross-validation and rigorous testing procedures to assess model performance under varying conditions.

   - Validation Metrics: Chose appropriate metrics that align with industry standards and stakeholder expectations, ensuring the reliability and usability of predictions.

4. **Integration with Operational Systems:**

   - Deployment Readiness: Designed models with deployment considerations in mind, ensuring compatibility with existing operational systems and scalability for future integration.

   - User-Friendly Interfaces: Developed user-friendly interfaces and documentation to facilitate adoption and usage by operational staff, minimizing resistance to change.

Evaluating Test Results Around Constraints: -

1. **Performance Metrics:**
    - Accuracy and Precision: Test results would typically include metrics such as accuracy, precision, recall, and F1-score for classification tasks, or mean squared error (MSE) and R-squared for regression tasks.
    - Impact of Constraints: Results would highlight how well the model performs under resource constraints (e.g., limited computational power) and data constraints (e.g., incomplete or noisy data).
2. **Validation Techniques:**
    - Cross-Validation: Results from cross-validation procedures would show the consistency of model performance across different subsets of data, indicating robustness against data variability.
    - Bias-Variance Trade-off: Analysis of bias and variance in model predictions would assess how well the model generalizes while considering constraints like limited training data or computational resources.
3. **Scalability and Efficiency:**
    - Computational Efficiency: Test results would demonstrate how efficiently the model operates within given computational constraints, such as processing time or memory usage.
    - Scaling Performance: Evaluation would include scalability tests to show how the model's performance scales with increasing data volumes or computational resources, highlighting potential bottlenecks or limitations.
4. **Compliance and Practicality:**
    - Regulatory Compliance: Results would discuss how well the model adheres to regulatory requirements or environmental standards, demonstrating its suitability for deployment in real-world settings.
    - Operational Integration: Testing would assess the ease of integrating the model into existing operational systems, including user acceptance and adaptability to operational workflows.

Interpretation of Test Results: -

1. **Trade-offs and Recommendations:**
   - Based on test results, recommendations would be made regarding trade-offs between model complexity and performance, considering resource constraints.
   - Insights would be provided on adjustments or optimizations needed to enhance model robustness and scalability while meeting industry standards and operational requirements.
2. **Decision Support:**
   - Test results would serve as decision support for stakeholders, providing clear insights into the feasibility and effectiveness of deploying the model in real industries.
   - Recommendations might include further iterations, adjustments to data collection processes, or enhancements to computational infrastructure to improve model performance under constraints.
3. **Future Directions:**
   - Results would inform future research or development directions, suggesting areas for improvement in handling constraints like data quality, scalability, or regulatory compliance.
   - Insights gained from test results would guide the refinement of methodologies, algorithms, and deployment strategies to better align with industry needs and constraints.

5.1 Test Plan/ Test Cases

Test Case 1: -

To ensure the predictive model for % Silica Concentrate meets performance, accuracy, and reliability requirements for real-time, minute-by-minute predictions.

Test Case 2: -

Build a predictive model to forecast % silica in concentrate several hours ahead to enable engineers to take proactive measures, optimizing processes and minimizing the percentage of iron lost to tailings.

Test Case 3: -

Determine if it is feasible to predict the percentage of silica in concentrate without using the % iron concentrate column, given their high correlation.

5.2 Test Procedure: -

The test procedure outlines the detailed steps required to execute the test plan and ensure that all test cases are systematically evaluated. Here's a comprehensive guide to the test procedure:

1. Preparation Phase:

1.1. Define Test Environment:

- Set up the necessary hardware and software environment for testing.
- Ensure that all required tools, libraries, and datasets are available and configured.

1.2. Prepare Test Data:

- Split the dataset into training, validation, and test sets.
- Ensure data integrity and quality for all splits.

2. Execution Phase:

2.1. Data Preprocessing Validation:

- Step 1: Load the raw dataset into the testing environment.
- Step 2: Apply the data preprocessing pipeline.
- Step 3: Check for remaining missing values and verify data normalization and scaling.
- Expected Outcome: Dataset is clean with no missing values and features are properly scaled.

2.2. Feature Engineering and Selection:

- Step 1: Apply feature engineering techniques to generate new features.
- Step 2: Calculate feature importance scores using techniques like correlation analysis or tree-based models.
- Step 3: Select top features based on importance scores.
- Expected Outcome: Relevant and high-quality features are selected for model training.

2.3. Model Training and Cross-Validation:

- Step 1: Split the dataset into training and validation sets using cross-validation.
- Step 2: Train the model on training folds.
- Step 3: Evaluate model performance on validation folds.
- Expected Outcome: Consistent performance across all folds with minimal variance.

2.4. Hyperparameter Tuning:

- Step 1: Define the hyperparameter search space.
- Step 2: Use grid search or random search to explore hyperparameter combinations.
- Step 3: Train and evaluate the model for each combination.
- Step 4: Select the best hyperparameters based on performance metrics.
- Expected Outcome: Optimal hyperparameters leading to improved model performance.

2.5. Model Evaluation on Test Set:

- Step 1: Load the test dataset.
- Step 2: Apply the trained model to the test dataset.
- Step 3: Evaluate performance using relevant metrics (e.g., accuracy, MSE, R-squared).
- Expected Outcome: Model performs well on the test set, meeting predefined thresholds.

2.6. Scalability Testing:

- Step 1: Simulate large datasets or use available large datasets.

- Step 2: Train and evaluate the model on the large dataset.
- Step 3: Monitor computational resource usage (e.g., CPU, memory).
- Expected Outcome: Model maintains performance and resource usage is within acceptable limits.

2.7. Integration and Deployment Testing:

- Step 1: Deploy the model in a test environment.
- Step 2: Integrate the model with existing systems and APIs.
- Step 3: Test model inference with real-time data inputs.
- Expected Outcome: Seamless integration, stable performance, and fast response times.

2.8. Compliance and Regulatory Testing:

- Step 1: Review relevant regulatory requirements and standards.
- Step 2: Validate data handling and model processes against these requirements.
- Step 3: Conduct audits and document compliance.
- Expected Outcome: Model and processes comply with all relevant regulations and standards.

3. Reporting Phase: -

3.1. Capture Test Results:

- Collect and log results from each test case.
- Document any anomalies or issues encountered during testing.

3.2. Analyze Results:

- Compare test results against expected outcomes.
- Identify areas of improvement or optimization.

3.3. Generate Test Report:

- Compile a comprehensive report detailing the test procedure, results, and any deviations.
- Include visualizations, charts, and graphs to illustrate key findings.

3.4. Review and Feedback:

- Share the test report with stakeholders for review.
- Collect feedback and incorporate it into future iterations of the project.

5.3 Performance Outcome: -

	Actual	Predicted
567416	2.08	2.097024
734100	3.26	3.317955
423833	1.29	1.382423
497626	2.46	2.786357
492026	1.45	1.590266
565417	2.08	2.090857
123578	5.25	5.155687
721990	1.82	2.280845
653037	4.25	4.255124
19018	5.30	5.349157
488978	2.08	2.049661
217065	1.34	1.430785
696214	1.47	1.675993
96318	3.70	3.594240
707093	1.00	1.244125
298621	1.13	1.193886
351547	2.10	1.860128
223674	1.86	1.845429
299349	1.19	1.255655
282110	1.07	1.183732

 # Test Case 1: -

Aim: To ensure the predictive model for % Silica Concentrate meets performance, accuracy, and reliability requirements for real-time, minute-by-minute predictions

Output: It is possible to predict % Silica Concentrate every minute by this model. Predicting % Silica Concentrate every minute in real-time would typically require a real-time data acquisition and prediction system.

Test Case 2: -

Aim: Build a predictive model to forecast % silica in concentrate several hours ahead to enable engineers to take proactive measures, optimizing processes and minimizing the percentage of iron lost to tailings

Output: It is possible to predict % Silica Concentrate ahead several hours by real time analysis of the data. But my model can't able show how many hours before it can predict.

The number of steps (hours) ahead you can predict % Silica in Concentrate depends on several factors:

1. **Data Granularity**: If your data is collected at hourly intervals, predicting several hours ahead (e.g., 1, 2, 3 hours) is more straightforward compared to predicting days or weeks ahead.
2. **Model Complexity**: The complexity of your predictive model and its ability to capture temporal dependencies in the data influence how far ahead you can reasonably predict. Models like autoregressive models (AR), autoregressive integrated moving average models (ARIMA), or recurrent neural networks (RNNs) can handle short-term predictions effectively.
3. **Data Quality and Stationarity**: The stationarity of your data (whether statistical properties remain constant over time) and the quality of your data (noise levels, missing values, outliers) affect prediction accuracy over longer time horizons.
4. **Domain Knowledge**: Understanding the mining process and how % Silica in Concentrate behaves over time can help in selecting appropriate features and models for longer-term predictions.

Test Case 3: -

Aim: Determine if it is feasible to predict the percentage of silica in concentrate without using the % iron concentrate column, given their high correlation

Output: Yes, it is possible to predict the % Silica in Concentrate without using the % Iron Concentrate column, but the accuracy of your prediction will depend on the availability and quality of other predictor variables in your dataset. You can use other features that might be correlated with % Silica in Concentrate.

6 My learnings: -

 # Summary of Overall Learning: -

1. **Technical Skills:**

Data Analysis and Machine Learning:

- Gained proficiency in data analysis techniques, including data cleaning, feature engineering, and statistical analysis.
- Learned to implement various machine learning algorithms for predictive modelling, such as regression, classification, and clustering.

Data Fusion and Feature Engineering:

- Learned methods for effective data fusion from diverse sources to enhance the robustness of prediction models.
- Developed skills in automated feature engineering and leveraging domain knowledge for creating relevant features.

Uncertainty Quantification and Model Explainability:

- Understood the importance of quantifying and reducing prediction uncertainty.
- Acquired knowledge of model explainability tools and techniques to ensure stakeholder trust and understanding.

Scalability and Deployment:

- Addressed scalability challenges to handle large-scale data and deploy models efficiently across various mining operations.
- Learned about deployment pipelines and integration with existing systems for seamless implementation.

2. **Soft Skills:**

Problem-Solving and Critical Thinking:

- Enhanced problem-solving skills by tackling complex challenges in quality prediction.

- Developed critical thinking abilities to evaluate different approaches and choose the best solutions.

Collaboration and Communication:

- Improved collaboration skills through teamwork and sharing insights with peers and stakeholders.
- Enhanced communication skills by presenting findings, explaining complex concepts, and writing comprehensive reports.

Project Management:

- Gained experience in managing projects, setting goals, and meeting deadlines.
- Learned to prioritize tasks and allocate resources efficiently.

How It Would Help in Career Growth: -

Industry Relevance:

- The acquired skills and knowledge are highly relevant to the mining industry, making me a valuable asset for companies focused on improving operational efficiency and quality control.
- Understanding advanced technologies and their applications positions me at the forefront of industry advancements.

Technical Expertise:

- The technical skills gained in data analysis and machine learning enhance my expertise and open up opportunities for roles such as Data Scientist, Machine Learning Engineer etc…

Versatility and Adaptability:

- The ability to handle diverse data sources, implement advanced models, and manage uncertainty equips me to tackle a wide range of problems across different domains.
- Adaptability to new tools and technologies ensures continuous growth and the ability to stay current with industry trends.

Leadership and Innovation:

- The experience gained in managing projects and collaborating with teams positions me for leadership roles where I can drive innovation and mentor others.
- Proposing future enhancements and demonstrating a forward-thinking approach showcases my potential to lead initiatives and contribute to the strategic direction of an organization.

Career Advancement:

- The combination of technical and soft skills enhances my qualifications for higher-level positions, increased responsibilities, and potential promotions.
- The ability to articulate and implement complex solutions makes me a strong candidate for roles that require both technical prowess and strategic vision.

7 Future work scope: -

Integration of Geospatial Data:

- Utilizing Geographic Information System (GIS) data to enhance prediction models by incorporating spatial variability and patterns.
- Developing methods to integrate remote sensing data, such as satellite imagery and LiDAR, for more comprehensive analysis.

Real-Time Data Processing:

- Implementing real-time data acquisition and processing systems to enable dynamic quality prediction and immediate decision-making.
- Exploring edge computing solutions to process data directly at mining sites for faster response times.

Enhanced Data Fusion Techniques:

- Combining data from various sources, including geological surveys, sensor data, historical production data, and environmental monitoring, to create more robust prediction models.
- Developing algorithms for effective data fusion and handling of heterogeneous data types.

Uncertainty Quantification:

- Incorporating techniques for quantifying and reducing prediction uncertainty to improve confidence in the models.
- Exploring probabilistic modelling approaches and Bayesian methods for better uncertainty management.

Automated Feature Engineering:

- Developing automated feature engineering techniques to identify and create relevant features from raw data, reducing manual intervention and improving model performance.
- Leveraging domain knowledge to create specialized features that capture important geological and operational characteristics.

Scalability and Deployment:

- Addressing scalability issues to ensure models can handle large-scale data and be deployed efficiently across different mining operations.
- Developing deployment pipelines and integration with existing mining software and systems for seamless implementation.

Environmental and Sustainability Considerations:

- Incorporating environmental impact assessments and sustainability metrics into the prediction models to promote eco-friendly mining practices.
- Exploring the use of prediction models to optimize resource extraction while minimizing environmental degradation.

Collaborative Platforms and Open Data Initiatives:

- Creating collaborative platforms for sharing data, models, and insights among different stakeholders in the mining industry.
- Promoting open data initiatives to facilitate research and innovation in quality prediction and other mining-related challenges.