Weekly Progress Report

Name: Md khurshid alam
Domain: Machine Learning
Date of submission:02/11/2025

Final Report: Predicting Silica

I. Overview:

This phase of the project focused on analyzing and predicting the % of Silica concentrate in iron ore from a real mining flotation plant dataset.

The objective was to build a time-series machine learning pipeline capable of forecasting silica impurities ahead of time so plant engineers can take corrective action.

Key steps included:

- Loading and cleaning a large industrial process dataset
- Handling mixed numeric formats and timestamp-based data
- Resampling to a uniform time frequency
- Engineering lag and rolling window features
- Applying regression models to predict future silica levels
- Evaluating whether silica can be predicted without using iron concentration

II. Achievements:

Data Processing & Cleaning

- Loaded dataset (~737k rows covering Mar–Sept 2017)
- Cleaned numeric fields with mixed comma formats (engineering-grade data)
- Converted timestamps to datetime and indexed by time
- Handled irregular sampling, resampled to 1-minute frequency
- Forward-filled hourly lab measurements for continuous signals

Feature Engineering

- Created lag features (1, 5, 10, 30, & 60-minute history)
- Added rolling mean & rolling standard deviation features
- Added time features (hour, minute, day of week)

Modeling & Evaluation

- Built a full forecasting pipeline with Random Forest
- Evaluated model for multi-step prediction horizons:
- 1, 2, 4, 8, 12, and 24 hours ahead
- Compared:
- Persistence baseline (last value carried forward)
- Random Forest model
- With & without % Iron Concentrate feature

Insights

- Model performed better than persistence baseline for short-term horizons
- Accuracy decreases as forecast horizon increases (as expected for industrial processes)
- Model was still able to provide actionable early warning signals
- Silica prediction is possible without Iron Concentrate feature, but accuracy is slightly reduced — confirming high correlation but also independence of predictive signal

III. Challenges

Challenge	Notes
Large dataset & memory handling	~737k rows needed efficient operations and time-based indexing
Mixed numeric formatting	Non-standard industrial formatting (comma decimal/thousands) required custom cleaning
Time-Series Nature	Needed lag, rolling windows, and careful train-test split
Error handling	Encountered & fixed version conflict in sklearn (squared parameter)
Engineering domain signals	Air flow, pulp density, levels needed understanding for meaningful features

IV. Learning Resources:

- Pandas documentation (time index, resampling)
- Scikit-learn documentation (regressors, metrics)
- YouTube tutorials on time-series ML & industrial data modeling
- Stack Overflow for comma-formatted numeric handling
- Matplotlib for visualization
- Research papers/blogs on flotation plant modeling and silica impurity prediction

V. Key Results & Insights:

- Short-term silica forecasting works effectively, allowing preventive actions
- Forecast horizon practicality: up to a few hours ahead is reliable
- % Iron Concentrate improves accuracy but model is usable without it → good for real-time systems where iron measurement may lag
- Validated importance of:
- Lagged sensor data
- Rolling process windows
- Real-world industrial predictive maintenance approach

VI. Conclusion:

This project successfully completed the end-to-end development of a real-world industrial ML pipeline using time-series mining plant data.

Major accomplishments include robust data cleaning, feature engineering, horizon forecasting, and evaluation of correlated features.
The results demonstrate the value of machine learning in predictive quality control for ore processing, supporting efficiency and environmental benefits.