**QUALITY PREDICTION IN MINING PROCESS**

**PROJECT PROPOSAL + INCREMENT REPORT**

**PROJECT PROPOSAL DESCRIPTION:**

**Project Title**: Quality prediction in mining process

**Team Members:**

Mounika Prathapani: 700745641

Srilaxmi Pavuluri: 700745445

Bharath Kumar Nakka: 700744145

**GOALS AND OBJECTIVES:**

**OBJECTIVE:**

Five main objectives of the project are:

* Building multiple machine learning regression models to predict the % of Silica Feed in Iron ore.
* Exploring the data variability.
* To reduce the variability in the data PCA dimensionality reduction method is applied.
* Consolidating the results of the models and conducting an experimental analysis
* Projecting PCA transformed data on to a 2-dimensional plane.

**SIGNIFICANCE:**

In general, assessing and predicting iron quality requires analyzing the complex relationship between variables, which demands domain expertise as well as chemical analysis, which is a time-consuming process. Datasets like mining contain highly correlated features because it is a combination of diverse factors like pulp quality and air flow. To mitigate the multicollinearity between the variables, dimensionality reduction method PCA is used along with the regression models.

To assess the model robustness various evaluation metrics are used.

**MOTIVATION:**

The percentage of Silica in Iron ore decides so many factors in real life, for example quality of the steel products. Silica is the most common impurity in iron ore extraction. The percentage of silica defines the purity of the iron. Knowing the accurate percentage of silica impurity helps to plan the manpower and operational costs. High amounts of silica form the slag and defines the quality of steel products in steel making the motivation of the project stems for the following advantages:

* In traditional laboratories the analysis takes a considerable amount of time. In machine learning the prediction and analyzing time is very less.
* In some situations, it becomes a tedious task to analyze a lot of parameters to predict the percentage of Silica.
* Machine learning quantifies the prediction results using a wide range of parameters.

**FEATURES:**

The main features of the project are:

* With accurate prediction of Silica, we can control the quality of a lot of products where silica is a main component, for example glass, steel, and ceramic.
* Another benefit of predicting the % of Silica is manpower planning and resource allocation in many industries.

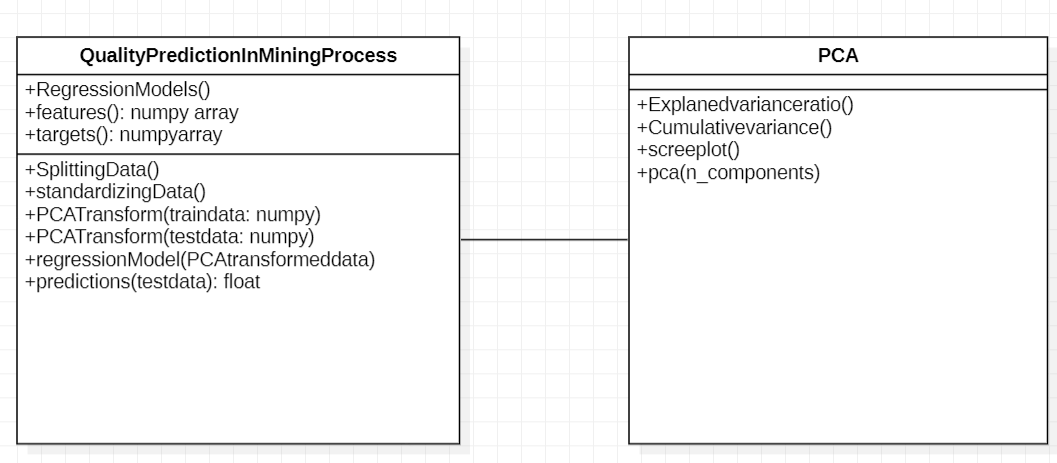
**DATASET:**

The dataset is collected from Kaggle open-source repository. Main aim is to predict the percentage of Silica that is impure in iron ore. The % of the silica column is sampled for every 20 seconds and few columns sampled on hourly basis. There is a timestamp for each row. Data is collected from March 2017 - September 2017.there are 737453 records and 24 features in the dataset.

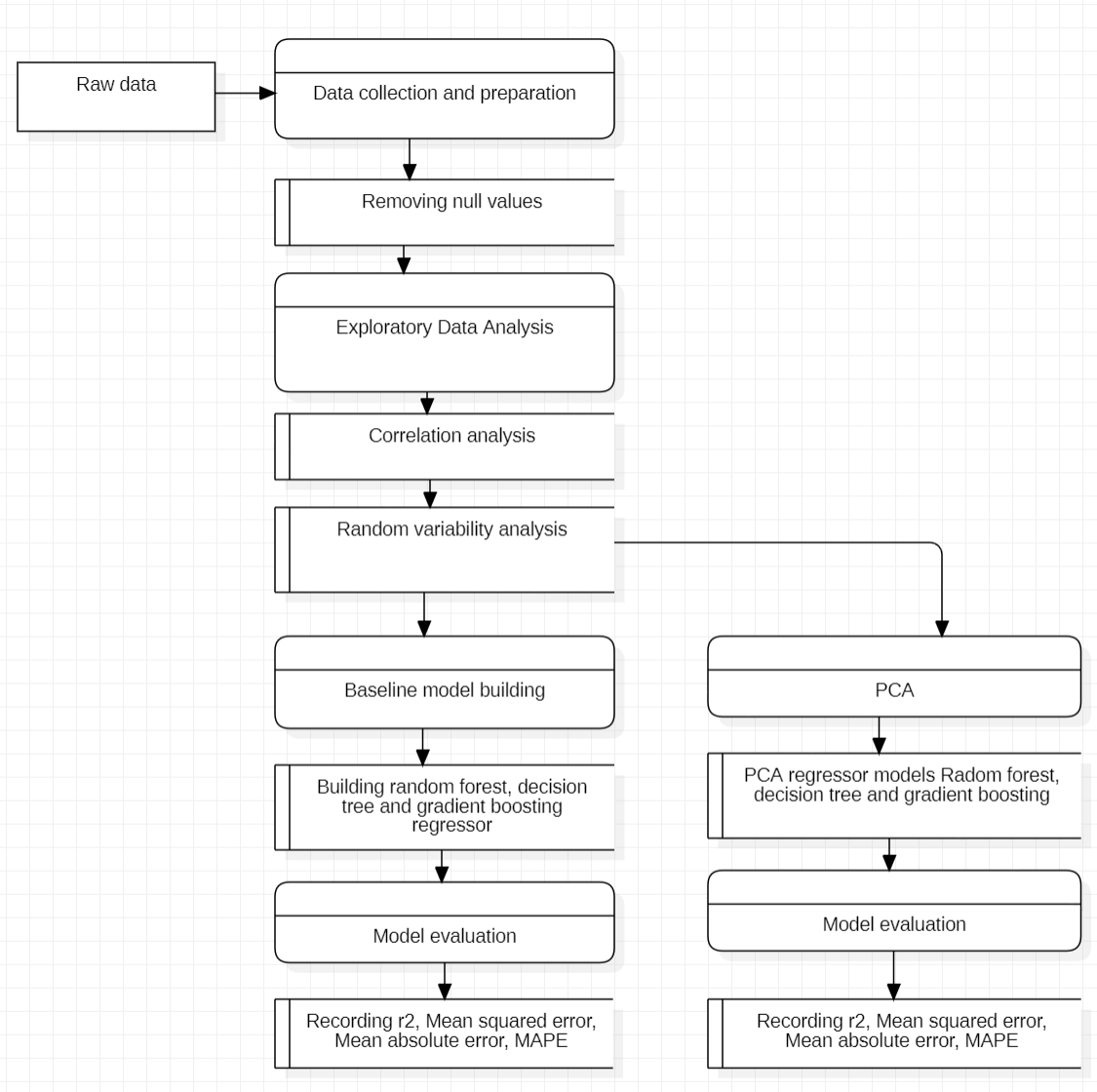
| Feature | Description |
| --- | --- |
| % Iron Feed | % of Iron that comes from the iron ore |
| % Silica Feed | % of silica (impurity) that comes from the iron ore |
| Starch Flow | Starch (reagent) Flow measured in m3/h |
| Amina Flow | Amina (reagent) Flow measured in m3/h |
| Ore Pulp Flow | t/h |
| Ore Pulp pH | pH scale from 0 to 14 |
| Ore Pulp Density | Density scale from 1 to 3 kg/cm³ |
| Flotation Column 02 Air Flow | Air flow that goes into the flotation cell measured in Nm³/h |

*table.1. Dataset description*

**DETAIL DESIGN OF FEATURES:**



*fig.1.Class diagram*



*fig.3.Data Flow Diagram*

**ANALYSIS:**

PCA is a technique for simplifying complex data. The PCA result has a simpler structure that may be processed using any algorithms, regression, or classification. We can project the high dimensional data onto a 2D plane. This study makes use of student academic performance data from Telkom University's Informatics Study Program and analyzes the student performance using time series analysis and classifies the performance using Naive Bayes algorithm [1].

Student learning activity data exhibits complexity. These Patterns can be utilized to increase the quality of learning. However, due to the complexity and high dimensionality of the data, information mining frequently encounters several challenges. Recently, numerous approaches for extracting meaningful information from complex data have been proposed; one potential method incorporates data visualization and clustering, which provides some level of insight into the data. This work describes the effectiveness of PCA on complex data [2].

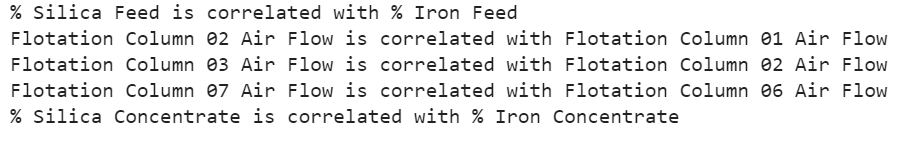
**IMPLEMENTATION:**

1.Data collection and preparation:

Raw data is collected and cleaned. In the cleaning process null values are removed. There are no null values in the data. In the preliminary stage statistical analysis of the data is observed using pandas’ data. Describe function to gain high level understanding of the features.

2.Exploratory Data Analysis:

In this step the relationship of the features is analyzed using various visualization techniques. In this step-in addition to the general exploration of data correlation analysis and random variability is checked. In the correlation analysis highly, correlated features are extracted using heatmap. There is a presence of multicollinearity in the data.



*fig.3.Correlation analysis report*

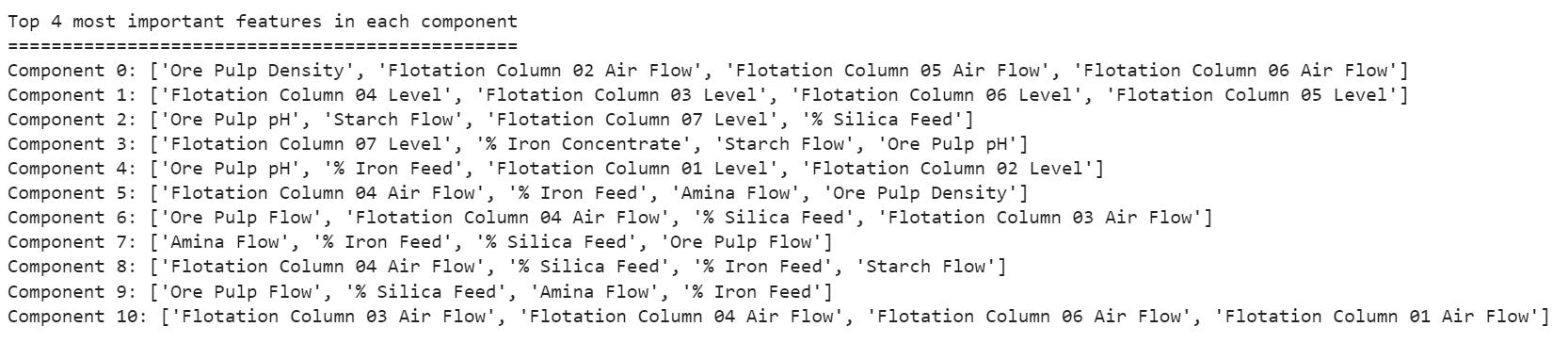
With time series analysis random variability of the data is observed *fig.6.*..There is a presence of random variability in the data.

3. Baseline models:

Baseline regression models are constructed using Random Forest regressor, Decision Tree regressor and Gradient Boosting regressor. Models are evaluated using various regression metrics r2 value, Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and MAPE.

4.Principal Component Analysis (PCA):

To withstand multicollinearity and random variability data is transformed using PCA. PCA object is initialized using scikit-learn module. Before we pass the data to the PCA object data is scaled using standard scaler. Scaled data is transformed using PCA. Cumulative variance and explained variance are visualized *fig.8,9*. From the visualization we observed that 85 of the variances is covered using 10 principal components. The PCA object is trained using 10 principal components. The top 4 most important features in each component are:

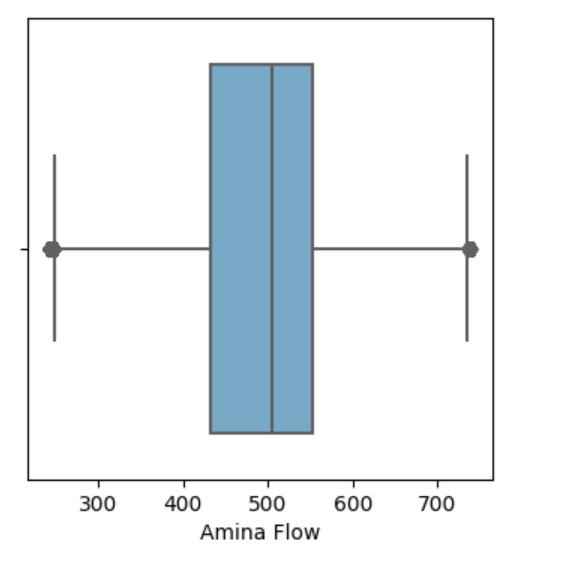


*fig.5.Top 4 most important features in each principal component*

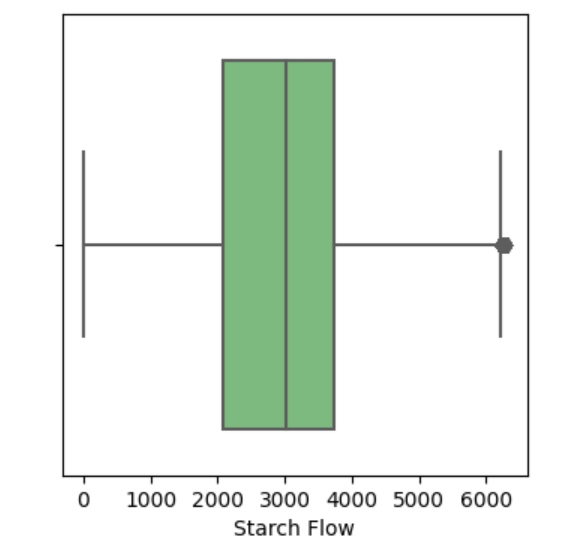
5.PCA regression models:

To check the impact of PCA on regression models. PCA transformed data is passed to the regression models and evaluated using regression metrics.

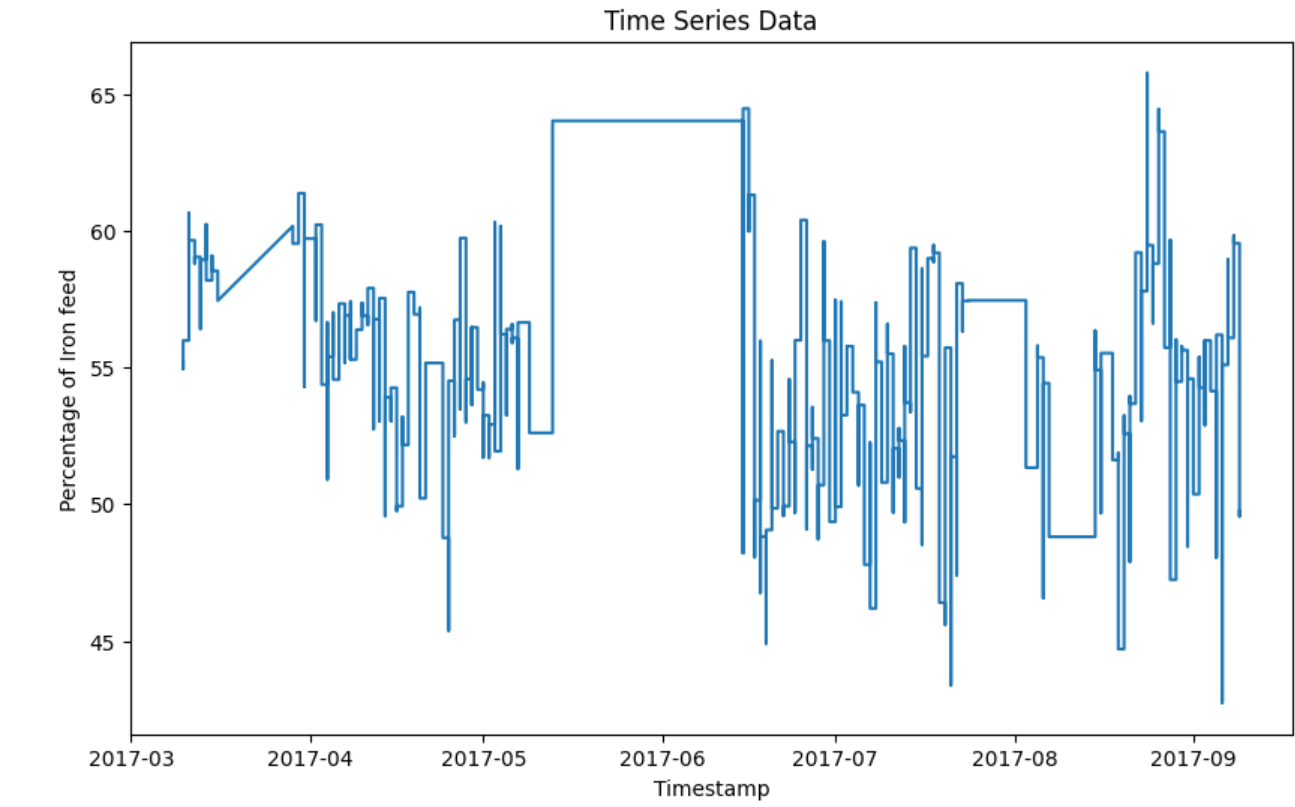
Preliminary Results:

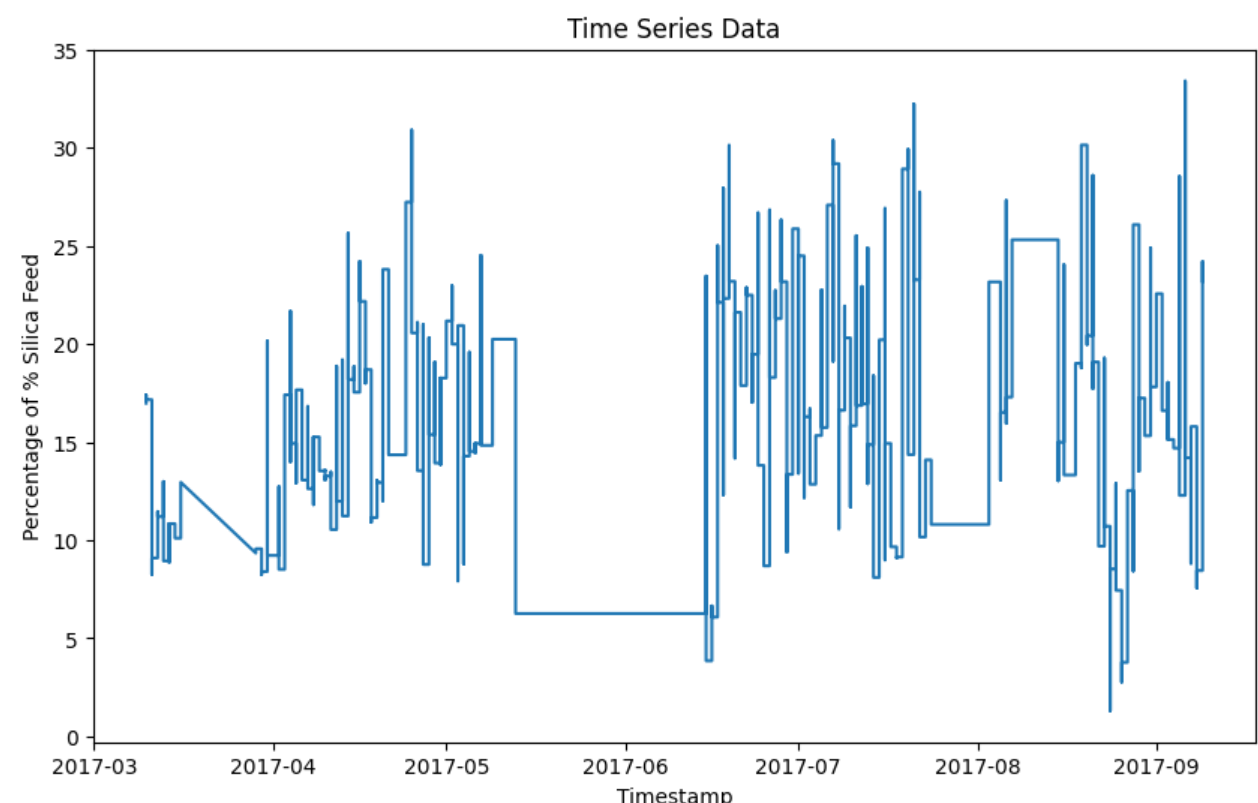


*fig.3.Distribution of Amina Flow*

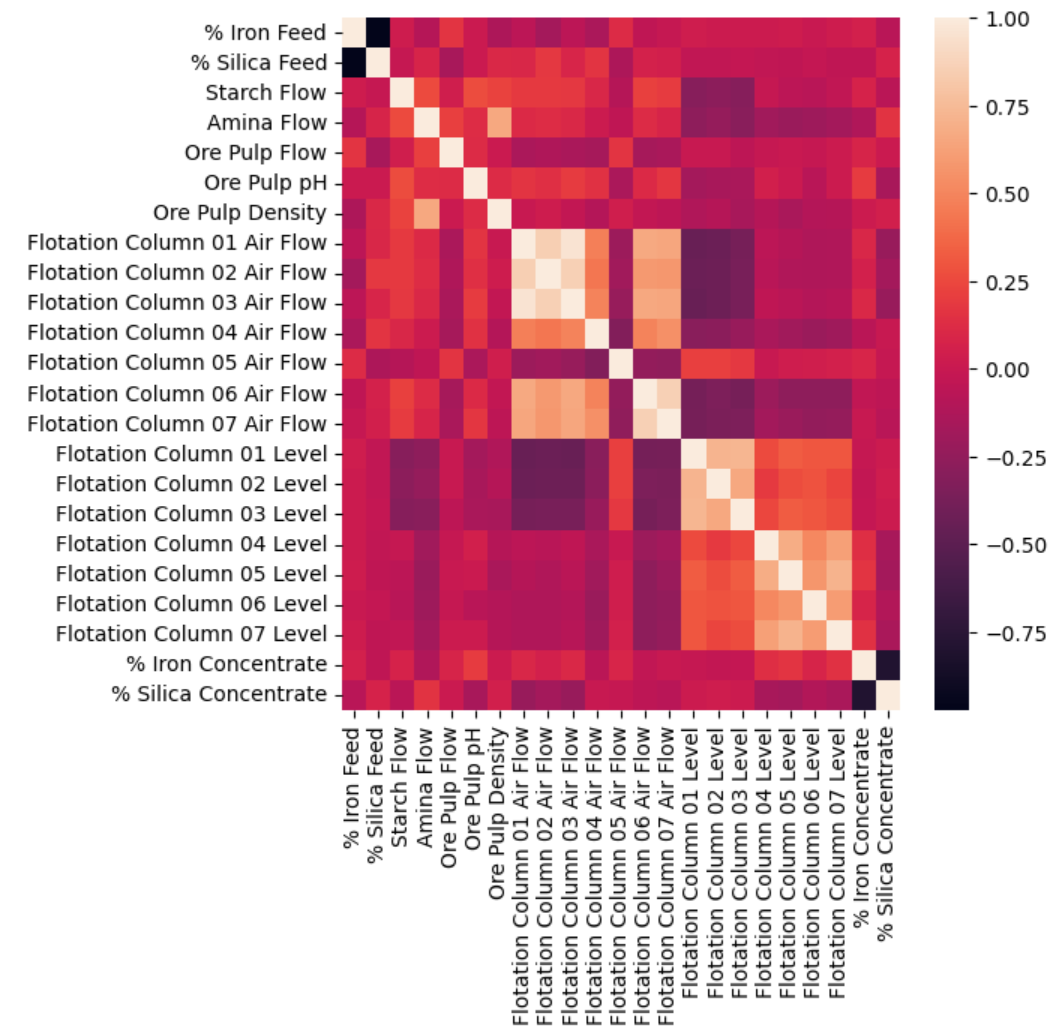
**

*fig.4.Starchflow distribution*

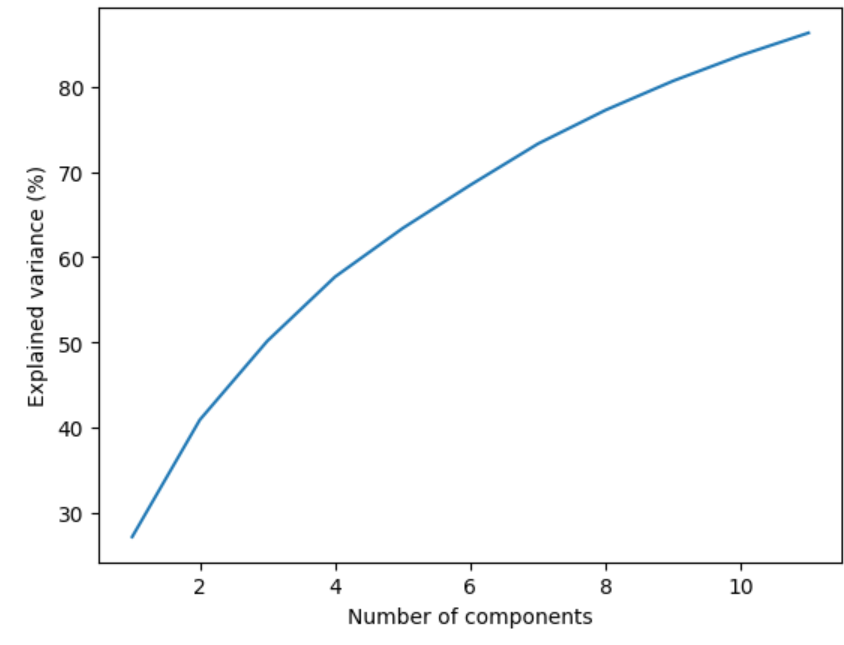
**

**

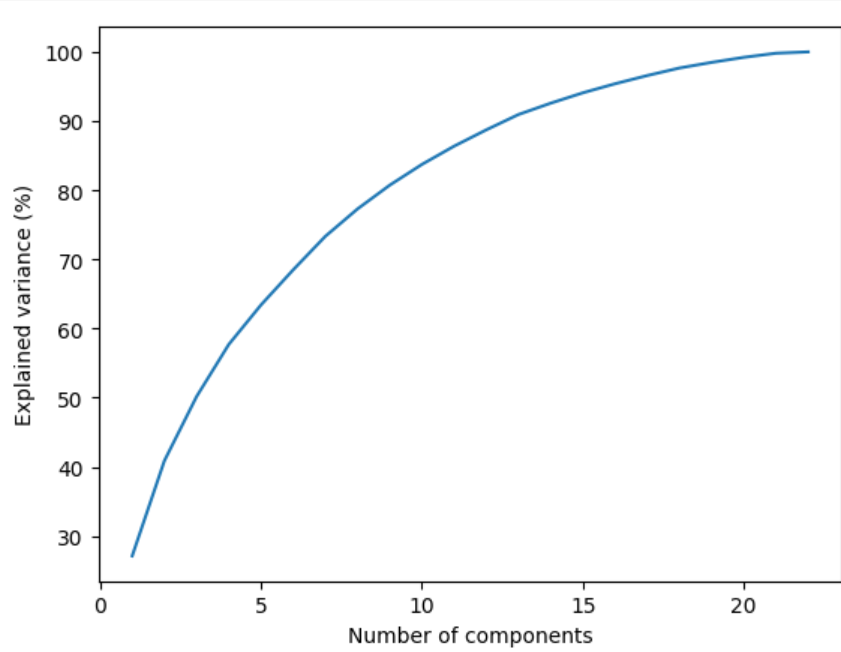
*fig.4.Time series analysis of Silica Feed and Iron feed*



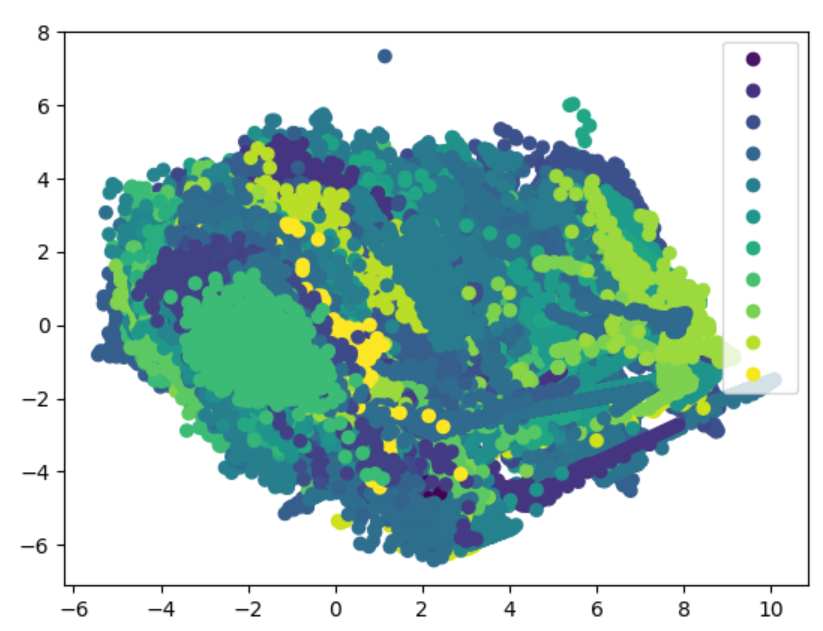
*fig.6.Correlation matrix*



*fig.3.Scree Plot*

**

*Fig.4.*

**

*fig.6.PCA transformed data.*

**PROJECT MANAGEMENT:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PROJECT MANAGEMENT REPORT | | | | |
| PROJECT NAME | Quality prediction in mining process | | | |
| REPORT DATE | 06/16/2023 | | | |
| PROJECT STATUS | Completed | | | |
| COMPLETED | 100% | | | |
|  | | | | |
| TASKS | DESCRIPTION | ASSIGNED TO | Contribution | STATUS |
| Data collection and preparation | Preprocessing of data | Bharath Kumar Nakka | 100% | COMPLETE |
| Exploratory Data Analysis | Exploring the data | Srilaxmi Pavuluri | 100% | COMPLETE |
| Baseline models | Building baseline models | Srilaxmi Pavuluri | 100% | COMPLETE |
| Principal Component Analysis | PCA | Mounika Prathapani | 100% | COMPLETE |
| Regression models | Building regression models on PCA transformed data | Mounika Prathapani | 100% | COMPLETE |
| Evaluation | Comparative analysis of results | Bharath Kumar Nakka | 100% | COMPLETE |
|  |  |  |  |  |

**REFERENCES:**

[1]. W. D. Herlambang, K. A. Laksitowening and I. Asror, "Prediction of Graduation with Naïve Bayes Algorithm and Principal Component Analysis (PCA) on Time Series Data," 2021 9th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, 2021, pp. 645-649, doi: 10.1109/ICoICT52021.2021.9527443.

[2].A. A. Supianto, T. Y. Christyawan, M. Hafis, Y. Hayashi, T. Hirashima and N. Hasanah, "Feature Dimensionality Reduction for Visualization and Clustering on Learning Process Data," 2019 International Conference on Sustainable Information Engineering and Technology (SIET), Lombok, Indonesia, 2019, pp. 84-89, doi: 10.1109/SIET48054.2019.8986020.