QUALITY PREDICTION IN A MINING PROCESS

A Project Report

Submitted to the Faculty of Engineering of
JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING

By

P.PRAVEEN **(17481A05D8)**

N.CHANDRIKA **(17481A05C1)**

P.VIGNESH **(17481A05C9)**

M. SRIKANTH **(18485A0522)**

Under the guidance of

Mr. M.N.SATISH KUMAR, M.Tech,(Ph.D)
Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GUDLAVALLERU ENGINEERING COLLEGE
(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRIRAO KNOWLEDGE VILLAGE
GUDLAVALLERU – 521356
ANDHRA PRADESH
2020-2021

QUALITY PREDICTION IN A MINING PROCESS

A Project Report

Submitted to the Faculty of Engineering of
**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA,
KAKINADA**

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE AND ENGINEERING

By

P.PRAVEEN **(17481A05D8)**

N.CHANDRIKA **(17481A05C1)**

P.VIGNESH **(17481A05C9)**

M.SRIKANTH **(18485A0522)**

Under the guidance of

Mr. M.N.SATISH KUMAR, M.Tech,(Ph.D)
Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GUDLAVALLERU ENGINEERING COLLEGE
(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRIRAO KNOWLEDGE VILLAGE
GUDLAVALLERU – 521356
ANDHRA PRADESH
2020-2021

GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"QUALITY PREDICTION IN A MINING PROCESS"** is a bonafide record of work carried out by **PATURI PRAVEEN (17481A05D8), NAKKA CHANDRIKA (17481A05C1), PALASALA VIGNESH (17481A05C9), MUKKU SRIKANTH (18485A0522)** under the guidance and supervision of **Mr. M.N.SATISH KUMAR** in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada** during the academic year 2020-21.

Project Guide	**Head of the Department**
(Mr. M.N.SATISH KUMAR)	**(Dr. M.BABU RAO)**

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Mr. M.N.Satish Kumar, Assistant Professor,** Department of Computer Science and Engineering for his constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M.Babu Rao**, Head of the Department, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. G.V.S.N.R.V Prasad** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project in time.

Team members

P.Praveen	(17481A05D8)
N.Chandrika	(17481A05C1)
P.Vignesh	(17481A05C9)
M.Srikanth	(18485A0522)

INDEX

TITLE **PAGENO**

LIST OF FIGURES

LIST OF ABBREVATIONS

Abbrevation	**Explanation**
ANN	Artificial Neural Networks
RNN	Recurrent Neural Network
MLR	Multiple Linear Regression
ML	Machine Learning
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
pH	Potential of Hydrogen

ABSTRACT

The% of Silica is measured in a lab experiment it takes at least one hour for the process engineers to have this value. As this impurity is measured every hour and it takes a lot of time for a day and causes delay in the mining process. The environment is polluting while reducing the number of ore that goes to tailings as you reduce silica in the ore concentrate. The overall goal is to predict impurity in the ore concentrate in mining process. In this case impurity is specifically Silica concentrate. Silica concentrate is a measured variable but takes time to report results, thus reducing efficiency in the mining process. Being able to predict the silica content without stopping to test is the extended goal of this project. This appears to be a continuous batch process, where raw material is fed into a flotation system, processed, removed, and the process repeated.

The purpose is to evaluate the feasibility of using machine learning algorithms like Multiple Linear Regression, Random Forest and Decision tree to predict in real-time. And also, by using Deep Learning techniques like LSTM, we can predict the silica impurity in the ore in less time and help the engineers for early prediction and reduce the impurities. We also developed a web application to display the prediction. The web application is built by using flask framework and it is integrated with trained ML model and it 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.

CHAPTER-I

INTRODUCTION

1.1 INTRODUCTION:

The investigation of primordial drivers regarding iron ore quality recovery in froth flotation processing plant has been of great interest lately. As seen in the Financial Times, ore iron, the large raw material for steel production is more integral to the global economy than any other commodity, except perhaps soil. It has been proven that approximately 2.5-3.0 tons of iron ore tailings are discharged for every one ton of iron ore produced. Moreover, statistics show that there is more than 130 million tons of iron ore produced annually. This indicates that if for example the mine tailing dams contain an average of approximately 12% iron ore, there would be approximately 1.52 million tons of iron wasted each fiscal year.

However, the recent boom in machine learning techniques offers an alternative paradigm for illustrating the relationships between real-time percentage of iron ore or silica concentrate and its superfluous features, thereby necessitating diverse models compared to the conventional laboratory approaches. It is also established that by employing powerful Artificial Intelligence model classes, such as Fuzzy Logic, Deep Neural Network, and PCA etc. the control and prediction efficiency quality iron recovery are significantly ameliorated. Generally, froth flotation processing plant has a vital role to play in the portfolio management.

The flotation is a mining industry process used to increase the amount of iron in the extracted ore. Chemical reagents are added into the ore pulp in order to modify its electrostatic properties and allow the removal of impurities, mainly in the form of silica, as a froth that can be easily collected. The output of this process is the concentrated ore, with a higher amount of iron and a lower amount of silica than the ore fed into the flotation plant. During the process several measurements are made and collected, referring both to the quality of the raw material (in the form of the percentage of iron and silica) and to the operational parameters of the flotation equipment. One way to measure the quality of flotation is to obtain the percentage of silica in the concentrated ore, however this procedure is performed in the laboratory and requires a time of up to two hours to provide results.

The froth is collected and dried under controlled sunlight. Most importantly, the percentage of silica concentrate and iron ore concentrate are ascertained at this crucial phase of the plant processes. Parameters such as pulp, operation conditions, appropriate dosage of reagent, mineralogy determined the ore recovery and the fastidious of the froth flotation process. Therefore, enterprises often make concerted efforts to collect empirical data at each processing phase and store in their repository database for consumption. The motivation is to liberate gangue from precious mineral, given a raw material which takes place in the froth flotation processing plant.

In order to realize the on-line quality monitoring activity, the ability to predict the finished product quality from manufacturing operation condition is required. This ability can be enabled by providing a formulation or mathematical model which can relate the manufacturing operation condition to the product quality.

This model is called quality prediction model. Using quality prediction model, process engineers are able to monitor product quality level by evaluating the manufacturing operation. Recently, various data mining techniques have been employed to develop quality prediction model from manufacturing.

For example, clustering, classification, association rules and regression have been applied in various industries. These techniques were implemented in injection molding industry, semiconductor manufacturing, slider manufacturing, machining process, hard disk manufacturing, loudspeaker manufacturing, and food processing industry. Most of the prediction models were developed in Single-stage Manufacturing System.

In the iron ore mining fraternity, in order to achieve the desired quality in the froth flotation processing plant, stakeholders rely on conventional laboratory test technique which usually takes more than two hours to ascertain.

Thus, the present study aims to evaluate the feasibility of using machine learning algorithms to predict the percentage of silica concentrate (SiO2) in the froth flotation processing plant in real-time. The selected features were then used in Multiple Linear Regression, Random Forest and Decision tree, LSTM models and the prediction accuracy of all the models have been evaluated and compared with each other.

Data storage in various format such as records, files, documents, sound, movies, science and a lot of new information forms has been driven by the emergence of information technology in numerous domains for better decision-making, the data generated from diverse applications require an appropriate way of extracting information from big repositories. The purpose is to discover usable information from the wide collection of data in databases (KDD), often dubbed data extraction (Data Mining). In order to find and extract patterns of recorded data, the basic features of data mining are numerous approaches and algorithms. Data mining and application for knowledge discovery have become an important component in many organizations, as they play an important part in decision making. In the new sectors of statistics, databases, machine learning, model reorganization and artificial intelligence, computer capabilities etc., data mining technology has been incorporated. In several industries, a model for quality prediction was established for the production of faultless products. However, in single-stage production, most quality prediction model is established. Previous research demonstrates that one-stage quality system in multi-stage production cannot effectively tackle the quality problem. Linear regression is a statistical study that determines how a relationship between two types of variables is modelled, dependent and independent (predictor).

Regression has major goal to investigate if independent factors have predicted the result variable well and which independent variables are important predictors of the result. There has recently been increasing interest in investigating primordial factors for iron ore recovery at a froth flotation treatment plant. The Financial Times shows that ore iron is more of a raw material for steel manufacture than any other commodity, with the exception perhaps of land, in the world economy. Every ton of iron ore produced has been shown to discharge around 2,5-3,0 tones. In addition, figures indicate that about 130 million tons of iron ore are produced annually.

This suggests that if the mining reservoirs contain, for instance, an average of around 12% iron ore, over 1.52 million tons of iron would be waste per year. In the brotherhood of iron mining, stakeholders rely on standard laboratory testing techniques, which generally take more than two hours to attain their target quality in the froth flotation processing facility. Since environmental protection is highly important and good iron grade is needed which is mined from ore. Then we may forecast a single dependent variable with two or three separate versions by applying machine learning methods such as Multiple Linear Regression, and we can employ Random Forest Decision Trees further. Deep learning techniques such as LSTM's neural network are utilized to predict silica impurity in mining process, which is well known for its time series prediction applications.

Considerations for applying data mining the user has to examine and clarify its purpose in order to develop a successful data mining solution. The problem target steers the user to the right learning algorithm paradigm. If the aim is to detect hidden groupings in the data or establish links between key data variables, users would like to discover information and choose a clustering or association mining algorithm. Alternatively, the aim could be to create a predictive model which can categories samples into a certain category such as low air quality or a real-world result like an air quality score. The prediction paradigm and the knowledge discovery paradigm are composed of a huge and increasing number of algorithms. It is an issue of its own accord to select amongst the procedures of any paradigm. Domingo's addresses some of the main aspects in helping practitioners who are new to the implementation of machine learning algorithms. The user should examine the complexity and amount of data provided while taking this decision. For instance, in a complex nonlinear classification task, a basic linear classifier is not useful. However, it demands users to take account of concerns relating to storage, memory and training time, and to use a substantial amount for advanced learning algorithms such as deep artificial neural networks.

Predicting quality prediction involves the development of models in which quality input features are related to quality inputs and the use of models in order to forecast what the resulting quality property value will be of a collection of input parameters. For predication, regression approach can be adjusted. The regression analysis can be used to model a connection among one or more independent and dependent variables. In the mining of data, independent variables are already known attributes, and we wish to anticipate the answer variables. Sadly, not just predictions are numerous real-world difficulties. Therefore, the prediction of future values can require more advanced algorithms (e.g. regression of logistics, decision trees or neural nets). For regression and classification, the same model types can often be utilized.

1.2 OBJECTIVES OF THE PROJECT:

- To evaluate the feasibility of using Machine Learning Algorithm like Multiple Linear Regression and Deep Learning Algorithm like LSTM to predict in real-time the percentage of silica concentrate of froth flotation processing plant.
- Predict silica percentage in an ore concentrate and report the results immediately without wasting the time in froth flotation processing plant.
- Estimate: The project will propose a model to predict percentage of silica concentrate in froth flotation.

- To develop a web application is built by using flask framework and it is integrated with trained ML model and it help the engineers, giving them early information to take actions (empowering!).

1.3 PROBLEM STATEMENT:

In practice, management, metallurgists and control operators, bank on laboratory test to take ad-hoc decision for corrective action in order to achieve optimal quality of ore recovery in the froth flotation system. Usually, the laboratory analysis takes two or more hours to ascertain the two variables of interest, which are the percentage of iron ore and silica concentrate. Such practice, however has demonstrated to be a non-novel technique to monitor and control the global unit circuit of the froth flotation system. In addition, the concomitant variation in the ore feed coupled with output stipulation changes make it cumbersome to put the plant in a steady state as a result of lengthy delays of the laboratory test.

This includes reduction of ore feed rate or increasing or decreasing the reagents flow or air flow. Although researchers have studied numerous different techniques on how to find a lasting solution to circumvent this relative efficient way of liberating gangue particles given a measurable input but little efforts have been made to estimate percentage of silica concentrate in real-time. More importantly, the silica concentrate is often quite susceptible to estimation error especially when it would end up in the tailings. Another problem worth mentioning is that company’s losses colossal amount of money each fiscal year as chucks of quality ore recovery end up in the tailings.

There is therefore the need to derive models for efficient and predictions that would incorporate all the stochastic nature of a flotation plant so as to build a reliable platform for decision making in real time.

CHAPTER – 2

LITERATURE REVIEW

Over the past two decades, there has been an upsurge of academic research work within froth flotation process fraternity. Though, a significant number of the plant processing problems are being successfully modelled using machine learning algorithms but other unresolved issues and impediment still remain.

Notably amongst papers which have presented similar work on using Artificial Intelligence in the froth flotation process plant worth mentioning is the work of Dawson [1] in which the researcher explored the use of Fuzzy Logic to control the flotation stochastic process, resulting in increased grade and recoveries.

The author further investigated the use of image analyzer coupled with Fuzzy logic and reagents control to provide a starting point for expert knowledge to be utilized in order to monitor, evaluate and control grade and recovery. He however averred that, despite the success, there are still opportunities for further enhancements within the control jurisdiction of employing image analyzing in the froth concentrate. The research concluded that availability of measurable features would be the basis for machine vision model to improve the froth flotation system problem.

Another author in the paradigm of froth flotation system has explored and discovered the most significant parameters that influence the flotation performance of lead mineral. Seemingly, the new model suggested that grinding time, flotation pH, for comparable collector, solid-in-pulp concentration and the increase of solid-in-pulp concentration have the most significant effect on the ore recovery and selective separation of lead mineral. He concluded that solid-in-pulp concentration was the most important parameter that influences the flotation of lead mineral [2].

In a recent study, advanced imaging systems based on Convolutional Neural Networks were employed to extract features from forth flotation plant, a case study of platinum flotation images at four distinct-grades. The extracted features were trained and compare with traditional texture feature extraction method.

It was found that the results were competitive and nearly comparable [3]. Another study examined the used of several Neural Networks architecture models to predict metallurgical performance of the flotation column at the Sarcheshmeh copper complex pilot plant. Apparently, 8 parameters were used namely: The chemical reagents dosage, froth height, air, wash water flow rates, gas holdup, Cu grades in the rougher feed, column feed and final concentrate streams concentrate streams were used for the simulation.

The authors proposed Artificial Neural Networks (ANN) and Multivariate 6 Nonlinear Regression (MNLR) as the most robust model for predicting copper ore recovery and grade [4]. However, Bergh el [5] all conducted a pilot and industrial research to establish how the characteristics of flotation processes, the quality of measurements of key variables and general lack of robust models are thwarting the appropriate use of predictive control. The authors proposed a multivariate statistics model such as PCA to explain the relationship between operation data for on-line diagnosis and fault detection. It was found that statistical methods seem to provide a general framework to build models in latent variables related to froth characteristics at short sampling internal.

The authors further discovered a supervisory and stabilizing control is a form of sub-optimal expert systems. The only challenge found was the difficulties in replicating a particular solution from one plant to another. Furthermore, another study proposed artificial neural networks as most robust predictive model to estimate the percentage of silica concentration in iron ore mining plant with the aid of virtual sensor. In his work, 700K observations and 120 parameters including desliming variables which correspond to 10 seconds sampling of the plant process and laboratory variables [6].

The above review highlights the myriad of issues informing and contextualizing the novel approach in the froth flotation system. From this overview, it is apparent that predicting real-time percentage of silica concentrate would be an essential component of the roll-out process. The chief objective, therefore in this thesis is to focus on using machine learning to predict online percentage of silica concentrate in the froth flotation process plant. More importantly, how to incorporate lagged values of silica concentration into the model.

CHAPTER – 3

PROPOSED METHOD

3.1 METHODOLOGY:

In this project we used one algorithm that is Multiple Linear Regression to build the model and also evaluate the dataset using Random Forest, Decision Tree, LSTM algorithms. Regression is used to predict silica impurity in an iron ore.

Regression:

Multiple linear regression, also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the independent variables and dependent variable.

A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

Multi Regression analysis is an extension of simple linear regression. It’s useful for describing and making predictions based on linear relationships between predictor variables (i.e., independent variables) and a response variable (i.e., dependent variable). Although multiple regression analysis is simpler than many other types statistical modelling methods, there are still some crucial steps that must be taken to ensure the validity of the results you obtain.

Before getting into any of the model investigations, make inspect and prepare your data. Check it for errors, treat any missing values, and inspect outliers to determine their validity.

In reality, there are multiple factors that predict the outcome of an event. The price movement of Exon Mobile, for example, depends on more than just the performance of the overall market. Other predictors such as the price of oil, interest rates, and the price scale of movement of oil futures can affect the price of XOM and stock prices of other oil companies. To understand a relationship in which more than two variables are present, a multiple linear regression is used.

Multiple linear regression (MLR) is used to determine a mathematical relationship among a number of random variables. In other terms, MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.

Regression residuals must be normally distributed. A linear relationship is assumed between the dependent variable and the independent variables. The residuals are homoscedastic and approximately rectangular-shaped.

At the center of the multiple linear regression analysis is the task of fitting a single line through a scatter plot. More specifically the multiple linear regression fits a line through a multi-dimensional space of data points. The simplest form has one dependent and two independent variables. The dependent variable may also be referred to as the outcome variable. The independent variables may also be referred to as the predictor variables or regressors.

There are 3 major uses for multi linear regression analysis:

First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable.

Second, it can be used to forecast effects or impacts of changes. That is, multiple linear regression analysis helps us to understand how much will the dependent variable change when we change the independent variables.

For instance, a multiple linear regression can tell you how much GPA is expected to increase (or decrease) for every one point to increase (or decrease) in IQ. Third, multi linear regression analysis predicts trends and future values. The multiple linear regression analysis can be used to get point estimates.

When selecting the model for the multi linear regression analysis, another important consideration is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable. Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model. Multiple Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages, analysing the correlation and directionality of the data and estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model.

Multi linear regression model would take the form:

$yi=\beta0+\beta1xi1+\beta2xi2+...+\beta pxip+\epsilon$

- where, for $i=n$ observations:
- yi=dependent variable
- xi=explanatory variables
- $\beta0$=y-intercept (constant term)
- βp=slope coefficients for each explanatory variable
- ϵ=the model's error term (also known as the residuals)
- x1, x2, …, xk are the predictors in the multiple regression model.

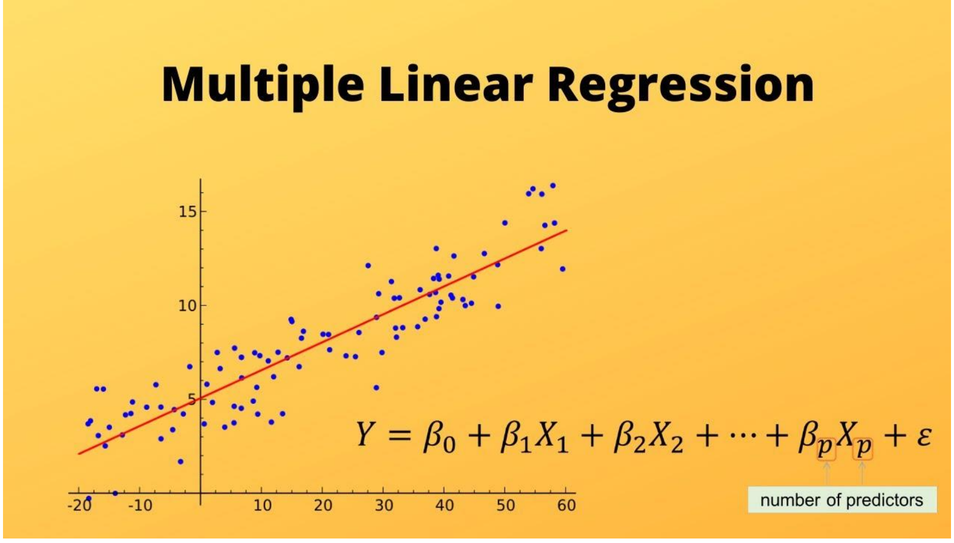


Fig 3.1.1 Multiple Linear Regression

Advantages of Multiple Linear Regression

- This model takes less time to predict the results.
- It gives accurate results.
- The algorithm used in this system easy to implement and it is also applicable to large datasets.
- Here we focus on groups of states with similar profiles so the result is accurate.

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is actually estimated.

Multiple linear regression can model more complex relationship which comes from various features together. They should be used in cases where one particular variable is not evident enough to map the relationship between the independent and the dependent variable.

When selecting the model for the multiple linear regression analysis, another important consideration is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable. Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model. Multiple Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points.

It consists of 3 stages analyzing the correlation and directionality of the data and estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model.

Random Forest:

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel.

There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta-estimator(i.e. it combines the result of multiple predictions) which aggregates many decision trees, with some helpful modifications.

The number of features that can be split on at each node is limited to some percentage of the total (which is known as the hyperparameter). This ensures that the ensemble model does not rely too heavily on any individual feature, and makes fair use of all potentially predictive features. Each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents overfitting.

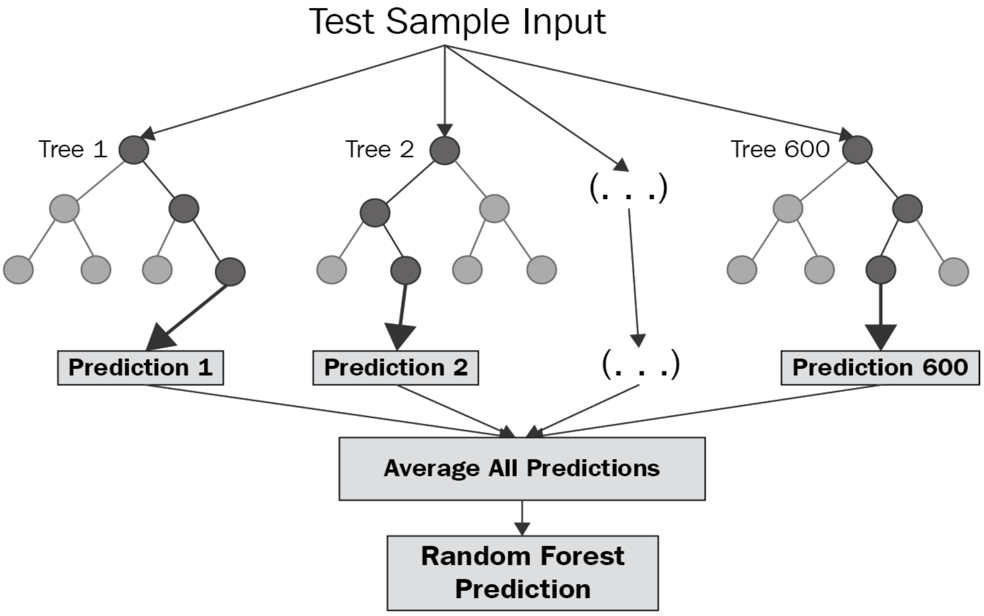


Fig 3.1.2 Random Forest

Decision Tree algorithm:

The decision tree model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, decision trees are able to capture non-linear interaction between the features and the target. Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node.

Disadvantages of Decision trees:

- Over fitting: Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning (discussed in detailed below).
- Not fit for continuous variables: While working with continuous numerical variables, decision tree loses information when it categorizes variables in different categories.
- Cannot extrapolate.
- Decision trees can be unstable: small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.
- No Guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.

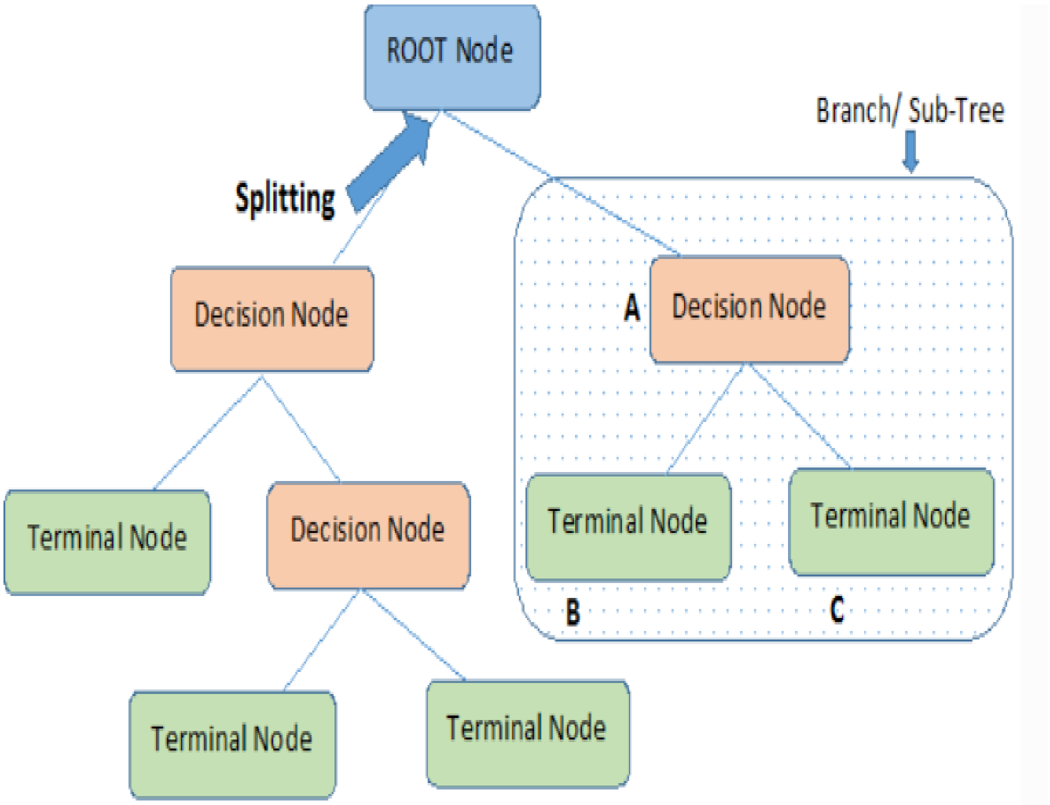


Fig 3.1.3 Decision Tree

Long Short-Term Memory:

LSTM is a recurrent neural network (RNN) architecture that REMEMBERS values over arbitrary intervals. LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, hidden Markov models and other sequence learning methods. The structure of RNN is very similar to hidden Markov model. However, the main difference is with how parameters are calculated and constructed. One of the advantages with LSTM is insensitivity to gap length. RNN and HMM rely on the hidden state before emission / sequence. If we want to predict the sequence after 1,000 intervals instead of 10, the model forgot the starting point by then. LSTM REMEMBERS.

The long-term memory is usually called the cell state. The looping arrows indicate recursive nature of the cell. This allows information from previous intervals to be stored with in the LSTM cell. Cell state is modified by the forget gate placed below the cell state and also adjust by the input modulation gate. From equation, the previous cell state forgets by multiply with the forget gate and adds new information through the output of the input gates.

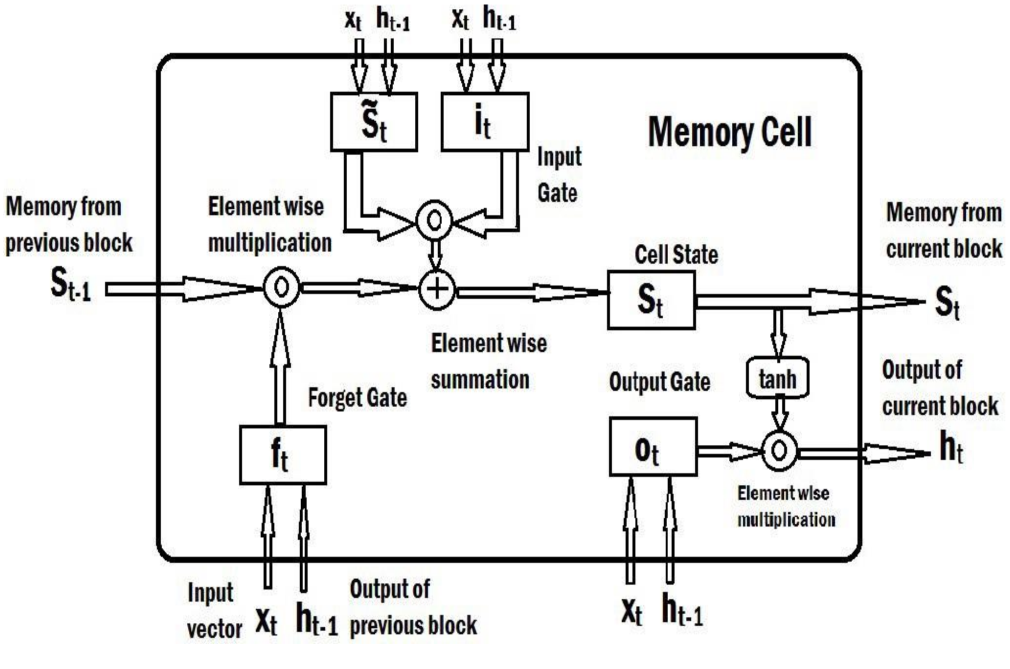

x_t h_{t-1} x_t h_{t-1}
$\tilde{S}_t$
i_t
Input Gate
Memory Cell
Memory from previous block
S_{t-1}
Element wise multiplication
Memory from current block
Cell State
S_t
S_t
Element wise summation
Output Gate
tanh
Output of current block
Forget Gate
f_t
o_t
h_t
Element wise multiplication
Input vector x_t h_{t-1} Output of previous block
x_t h_{t-1}

Fig 3.1.4 Basic Structure of LSTM

Flask Framework:

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies, and several common framework-related tools.

Flask is very easy to learn and quick to implement. It is preferred by many Data Scientists as it is quick to implement and easy to show a demo of their models. As Flask is written in Python, Developers who are familiar with Python find it no difficult to learn Flask. Though Flask has no additional tools or libraries that a normal framework would have, it makes it easy to implement other services like databases.

Flask is a lightweight Web Server Gateway Interface Web Application (WSGI) framework. It gives the developer varieties of choice when developing web applications, it provides you with the necessary tools to build and deploy a web application. It does not enforce any dependencies or give a fixed project structure like other Python Web Frameworks like Django offers. This adds to the advantage side by making it useable by many developers. The website can be on anything, but still, it allows the developers the opportunity to use some extensions provided by the community that allows you to add more functionality to the web application.As said above Flask is a lightweight Web Server Gateway Interface (WSGI) web application framework. It has minimal or no external libraries. It is used to create a simple website and then scale it up to complex applications. Below are some of the merits of using Flask.

Easy to Use: The Flask framework is easy to understand, which makes it perfect for beginners. The simplicity in the flask framework enables the developer to navigate around and create the application easily.

Very Flexible: Most of the Flask components can be altered. It allows users to customize the website.

Testing: It allows unit testing through its integrated support, built-in development server, fast debugger, and restful request dispatching. Makes testing much faster.

Project Structure:

This project has four parts:

model.py — This contains code for the machine learning model to predict sales in the third month based on the sales in the first two months.

app.py — This contains Flask APIs that receives sales details through GUI or API calls, computes the predicted value based on our model and returns it.

request.py — This uses requests module to call APIs defined in app.py and displays the returned value.

HTML/CSS — This contains the HTML template and CSS styling to allow user to enter sales detail and displays the predicted sales in the third month.

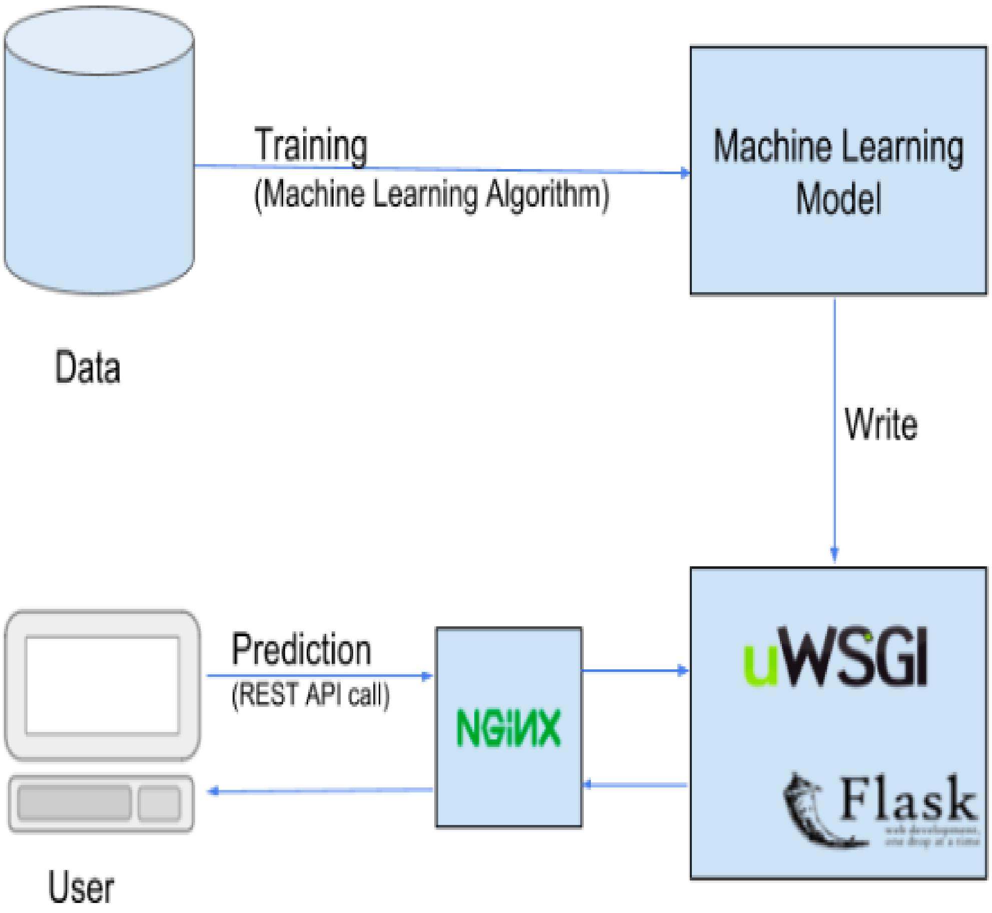
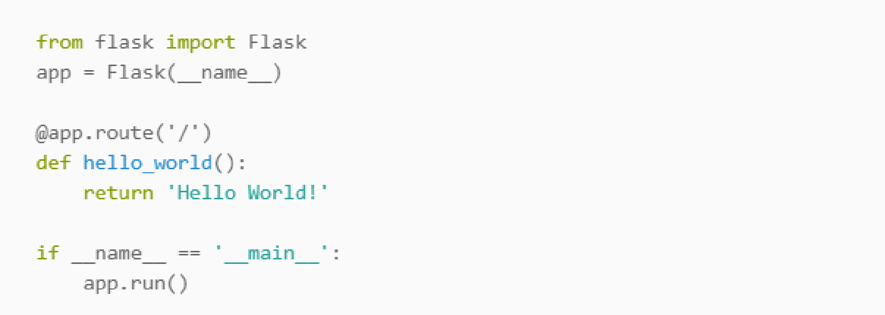


Fig 3.1.5 Deployment of a Machine Learning model

Unlike the Django framework, Flask is very Pythonic. It’s easy to get started with Flask, because it doesn’t have a huge learning curve.

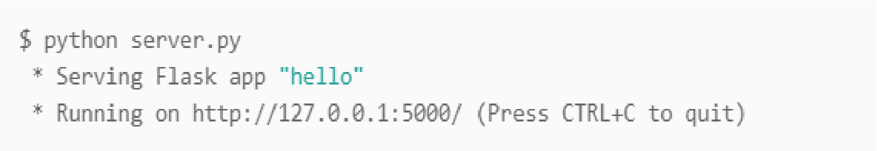
On top of that it’s very explicit, which increases readability. To create the “Hello World” app, you only need a few lines of code.

This is a boilerplate code example.



```
from flask import Flask
app = Flask(__name__)

@app.route('/')
def hello_world():
    return 'Hello World!'

if __name__ == '__main__':
    app.run()
```

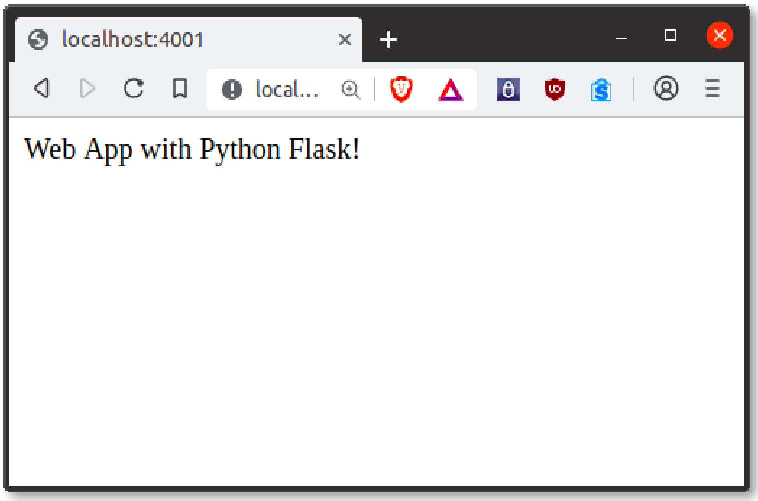
If you want to develop on your local computer, you can do so easily. Save this program as server.py and run it with python server.py.



```
$ python server.py
 * Serving Flask app "hello"
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

It then starts a web server which is available only on your computer. In a web browser open localhost on port 5000 (the url) and you’ll see “Hello World” show up. To host and develop online, you can use PythonAnywhere.

Some example output:



Web App with Python Flask!

Fig 3.1.6 Python Flask

3.2 IMPLEMENTATION:

In order to implement our model, we have to follow the following steps:

Step 1: Importing Libraries

We used pandas, numpy and Multilinear Regression classifier for prediction. Pandas will be used for performing operations on data frames. Further, more using numpy, we will perform necessary mathematical operations.

Numpy:

Numpy is a Python package which stands for Numerical Python. It is the core library for scientific computing, which contains a powerful n-dimensional array object, provide tools for integrating C, C++ etc. It is also useful in linear algebra, random number capability etc.

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with python. Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

Array in Numpy is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In Numpy, number of dimensions of the array is called rank of the array. A tuple of integers giving the size of the array along each dimension is known as shape of the array. An array class in Numpy is called as ndarray. Elements in Numpy arrays are accessed by using square brackets and can be initialized by using nested Python.

Pandas:

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib:

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB.

Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open-source. Several toolkits are available which extend Matplotlib functionality. Some are separate downloads; others ship with the Matplotlib source code but have external dependencies.

Visualization with Matplotlib- One of Matplotlib‛s most important features is its ability to play well with many operating systems and graphics backends. Matplotlib supports dozens of backends and output types, which means you can count on it to work regardless of which operating system you are using or which output format you wish. This cross-platform, everything-to-everyone approach has been one of the great strengths of Matplotlib. It has led to a large user base, which in turn has led to an active developer base and Matplotlib‛s powerful tools and ubiquity within the scientific Python world.

In recent years, however, the interface and style of Matplotlib have begun to show their age. Newer tools like ggplot and ggvis in the R language, along with web visualization toolkits based on D3js and HTML5 canvas, often make Matplotlib feel clunky and old-fashioned. Still, I'm of the opinion that we cannot ignore Matplotlib's strength as a well-tested, crossplatform graphics engine. Recent Matplotlib versions make it relatively easy to set new global plotting styles (see Customizing Matplotlib: Configurations and Style Sheets), and people have been developing new packages that build on its powerful internals to drive Matplotlib via cleaner, more modern APIs—for example, Seaborn (discussed in Visualization With Seaborn), ggpy, HoloViews, Altair, and even Pandas itself can be used as wrappers around Matplotlib's API. Importing Matplotlib - Just as we use the npshorthand for NumPy and the pd shorthand for Pandas, we will use some standard shorthands for Matplotlib imports.

Sklearn :

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

Seaborn:

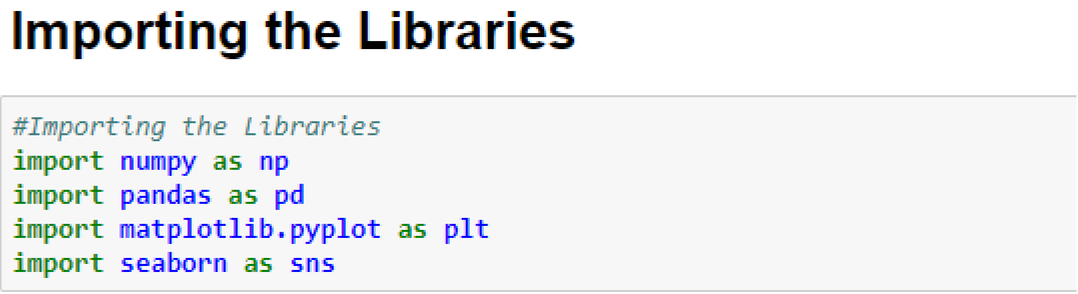
Seaborn is a wonderful visualization library provided by python. It has several kinds of plots through which it provides the amazing visualization capabilities. Some of them include count plot, scatter plot, pair plots, regression plots, matrix plots and much more.

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Here is some of the functionality that seaborn offers:

- A dataset-oriented API for examining relationships between multiple variables
- Specialized support for using categorical variables to show observations or aggregate statistics
- Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data
- Automatic estimation and plotting of linear regression models for different kinds dependent variables
- Convenient views onto the overall structure of complex datasets
- High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
- Concise control over matplotlib figure styling with several built-in themes
- Tools for choosing color palettes that faithfully reveal patterns in your data

Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.



Importing the Libraries

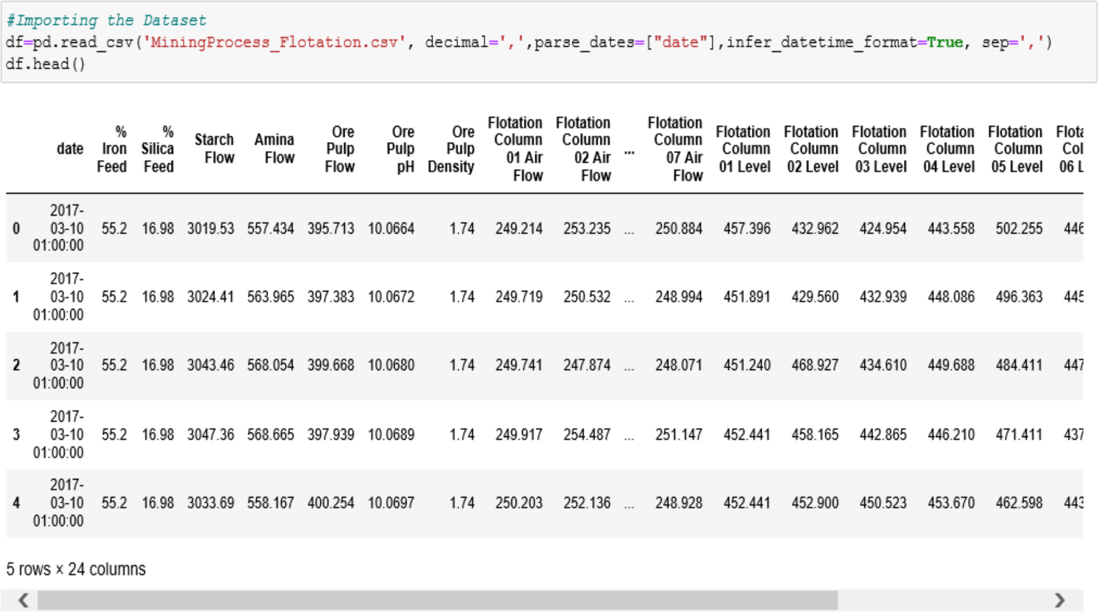
```
#Importing the Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Fig 3.2.1 Importing Libraries

Step 2: Reading the dataset

To pick the right variables, you have got to have a basic understanding of your dataset, enough to know that your data is relevant, high quality, and of adequate volume. As part of

your model building efforts, you will be working to select the best predictor variables for model.



```
#Importing the Dataset
df=pd.read_csv('MiningProcess_Flotation.csv', decimal=',',parse_dates=["date"],infer_datetime_format=True, sep=',')
df.head()
```

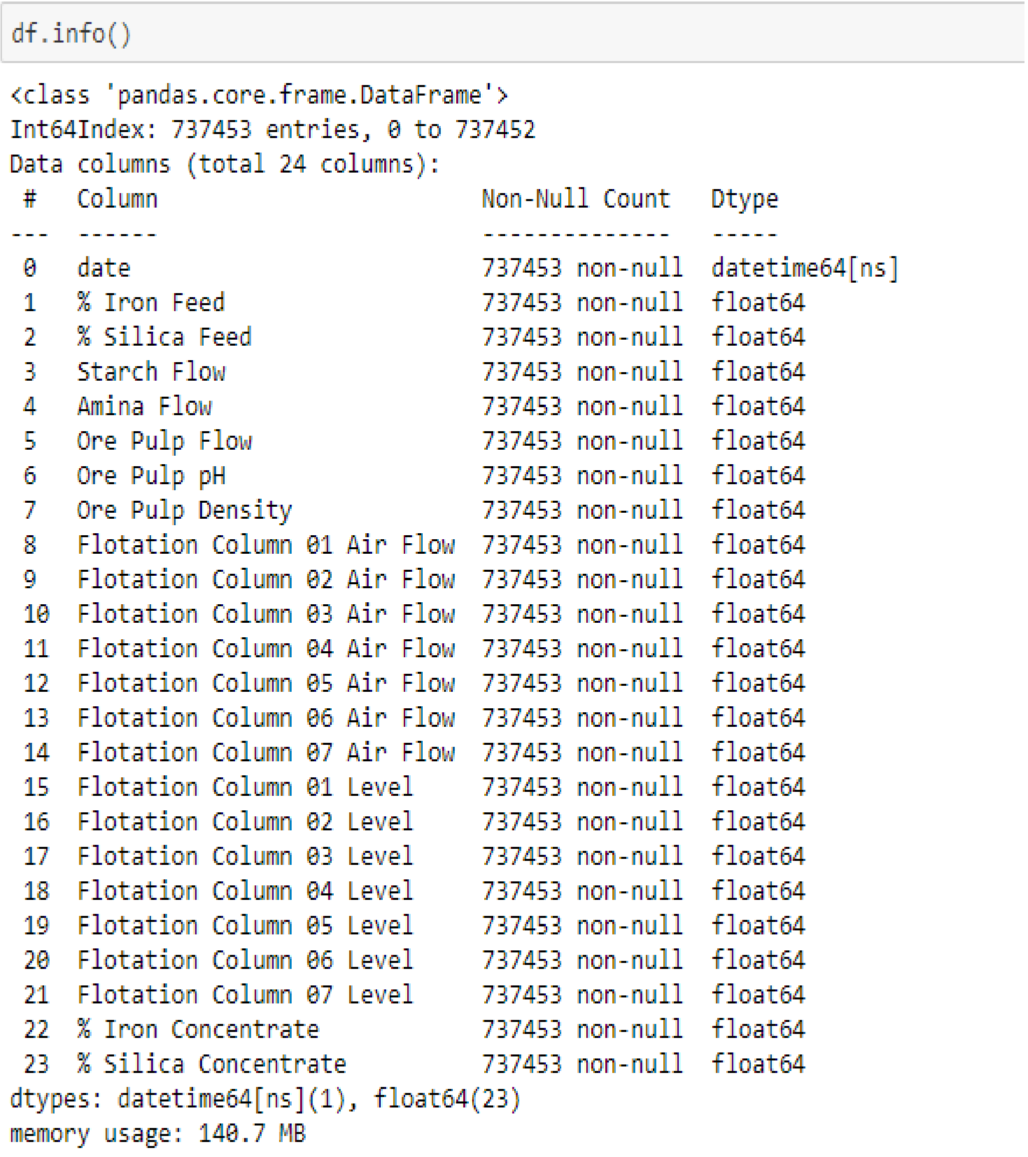
	date	% Iron Feed	% Silica Feed	Starch Flow	Amina Flow	Ore Pulp Flow	Ore Pulp pH	Ore Pulp Density	Flotation Column 01 Air Flow	Flotation Column 02 Air Flow	...	Flotation Column 07 Air Flow	Flotation Column 01 Level	Flotation Column 02 Level	Flotation Column 03 Level	Flotation Column 04 Level	Flotation Column 05 Level	Flota Col 06 L
0	2017-03-10 01:00:00	55.2	16.98	3019.53	557.434	395.713	10.0664	1.74	249.214	253.235	...	250.884	457.396	432.962	424.954	443.558	502.255	446
1	2017-03-10 01:00:00	55.2	16.98	3024.41	563.965	397.383	10.0672	1.74	249.719	250.532	...	248.994	451.891	429.560	432.939	448.086	496.363	445
2	2017-03-10 01:00:00	55.2	16.98	3043.46	568.054	399.668	10.0680	1.74	249.741	247.874	...	248.071	451.240	468.927	434.610	449.688	484.411	447
3	2017-03-10 01:00:00	55.2	16.98	3047.36	568.665	397.939	10.0689	1.74	249.917	254.487	...	251.147	452.441	458.165	442.865	446.210	471.411	437
4	2017-03-10 01:00:00	55.2	16.98	3033.69	558.167	400.254	10.0697	1.74	250.203	252.136	...	248.928	452.441	452.900	450.523	453.670	462.598	443

5 rows × 24 columns

Fig 3.2.2 Dataset

Step3: Checking for null values:

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously, you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 737453 entries, 0 to 737452
Data columns (total 24 columns):
 #   Column                        Non-Null Count   Dtype         
---  ------                        --------------   -----         
 0   date                          737453 non-null  datetime64[ns]
 1   % Iron Feed                   737453 non-null  float64       
 2   % Silica Feed                 737453 non-null  float64       
 3   Starch Flow                   737453 non-null  float64       
 4   Amina Flow                    737453 non-null  float64       
 5   Ore Pulp Flow                 737453 non-null  float64       
 6   Ore Pulp pH                   737453 non-null  float64       
 7   Ore Pulp Density              737453 non-null  float64       
 8   Flotation Column 01 Air Flow  737453 non-null  float64       
 9   Flotation Column 02 Air Flow  737453 non-null  float64       
 10  Flotation Column 03 Air Flow  737453 non-null  float64       
 11  Flotation Column 04 Air Flow  737453 non-null  float64       
 12  Flotation Column 05 Air Flow  737453 non-null  float64       
 13  Flotation Column 06 Air Flow  737453 non-null  float64       
 14  Flotation Column 07 Air Flow  737453 non-null  float64       
 15  Flotation Column 01 Level     737453 non-null  float64       
 16  Flotation Column 02 Level     737453 non-null  float64       
 17  Flotation Column 03 Level     737453 non-null  float64       
 18  Flotation Column 04 Level     737453 non-null  float64       
 19  Flotation Column 05 Level     737453 non-null  float64       
 20  Flotation Column 06 Level     737453 non-null  float64       
 21  Flotation Column 07 Level     737453 non-null  float64       
 22  % Iron Concentrate            737453 non-null  float64       
 23  % Silica Concentrate          737453 non-null  float64       
dtypes: datetime64[ns](1), float64(23)
memory usage: 140.7 MB
```

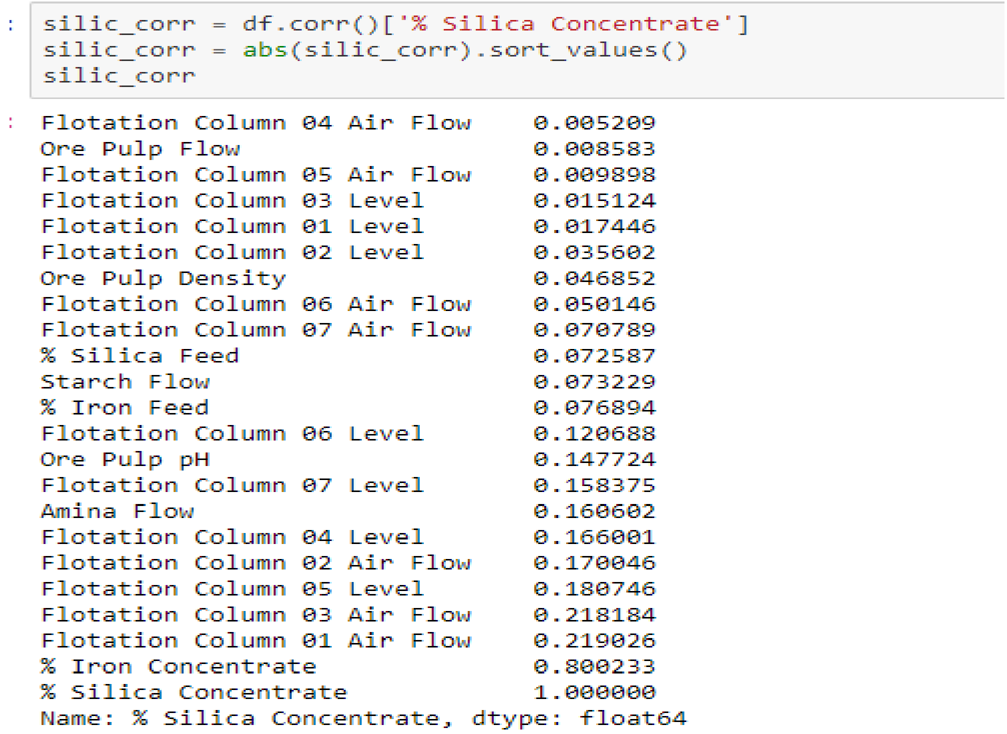
Fig 3.2.3 Dataset Information

Step4: Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. We have carried below preprocessing steps. In this model implementation we used interquartile range (IQR), it is the distance between the first and third quartile marks. The IQR is a measurement of the variability about the median. More specifically, the IQR tells us the range of the middle half of the data and removed 6 outliers.

Step5: Finding correlation coefficient

Pandas dataframe.corr() is used to find the pairwise correlation of all columns in the dataframe by using Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the-relationship. K



```
silic_corr = df.corr()['% Silica Concentrate']
silic_corr = abs(silic_corr).sort_values()
silic_corr
```

```
Flotation Column 04 Air Flow    0.005209
Ore Pulp Flow                   0.008583
Flotation Column 05 Air Flow    0.009898
Flotation Column 03 Level       0.015124
Flotation Column 01 Level       0.017446
Flotation Column 02 Level       0.035602
Ore Pulp Density                0.046852
Flotation Column 06 Air Flow    0.050146
Flotation Column 07 Air Flow    0.070789
% Silica Feed                   0.072587
Starch Flow                     0.073229
% Iron Feed                     0.076894
Flotation Column 06 Level       0.120688
Ore Pulp pH                     0.147724
Flotation Column 07 Level       0.158375
Amina Flow                      0.160602
Flotation Column 04 Level       0.166001
Flotation Column 02 Air Flow    0.170046
Flotation Column 05 Level       0.180746
Flotation Column 03 Air Flow    0.218184
Flotation Column 01 Air Flow    0.219026
% Iron Concentrate              0.800233
% Silica Concentrate            1.000000
Name: % Silica Concentrate, dtype: float64
```

Fig 3.2.4 Correlation Coefficients

Step6: Training and Test Sets: Splitting Data

The module introduced the idea of dividing your data set into two subsets:

training set—a subset to train a model.

Test set—a subset to test the trained model.

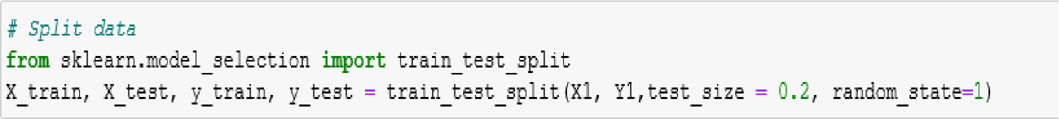
You could imagine slicing the single data set as Slicing a single data set into a training set and test set.

Make sure that your test set meets the following two conditions:

- Is large enough to yield statistically meaningful results.
- Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data. Our test set serves as a proxy for new data. For example, consider the following figure. Notice that the model learned for the training data is very simple. This model doesn't do a perfect job—a few predictions are wrong. However, this model does about as well on the test data as it does on the training data. In other words, this simple model does not overfit the training data.

We have splited the the dataset as 80% of the data to train the model and 20% to test the model. For this we need to import the "train_test_split" from sklearn package. In this splitting we use class which have attributes like test_size that specifies the percentage of test data and random_state that can have values 0 or 1 which is used to set the test data from the dataset.



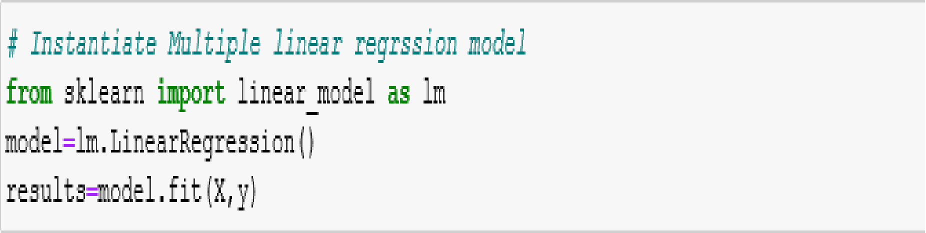
```
# Split data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X1, Y1,test_size = 0.2, random_state=1)
```

Fig 3.2.5 Splitting Data

In this the variables X_train, y_train refers to the training data and X_test, y_test refers to the test data. The train values are passed for training the model and the test values are passed for testing the model developed.

Step 7: Model Training

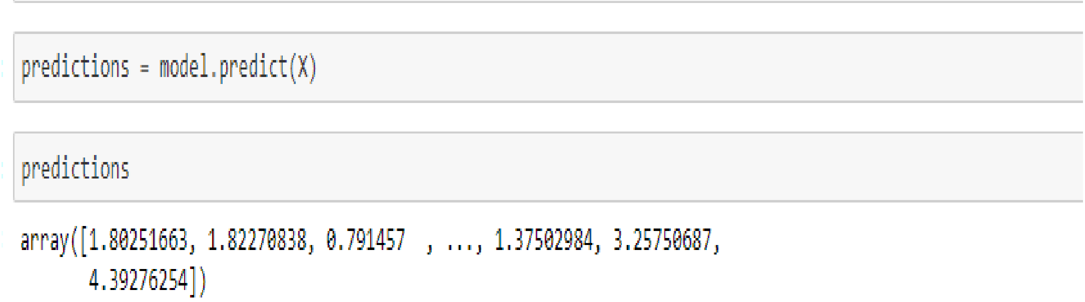
To train the model we need to import the required model. As we are using the Multiple Linear Regression we need to import the LinearRegression class form sklearn.linear_model library. And the model is trained by passing the trian data (x_train and y_train)



```python
# Instantiate Multiple linear regrssion model
from sklearn import linear_model as lm
model=lm.LinearRegression()
results=model.fit(X,y)
```

Fig 3.2.6 Model Training

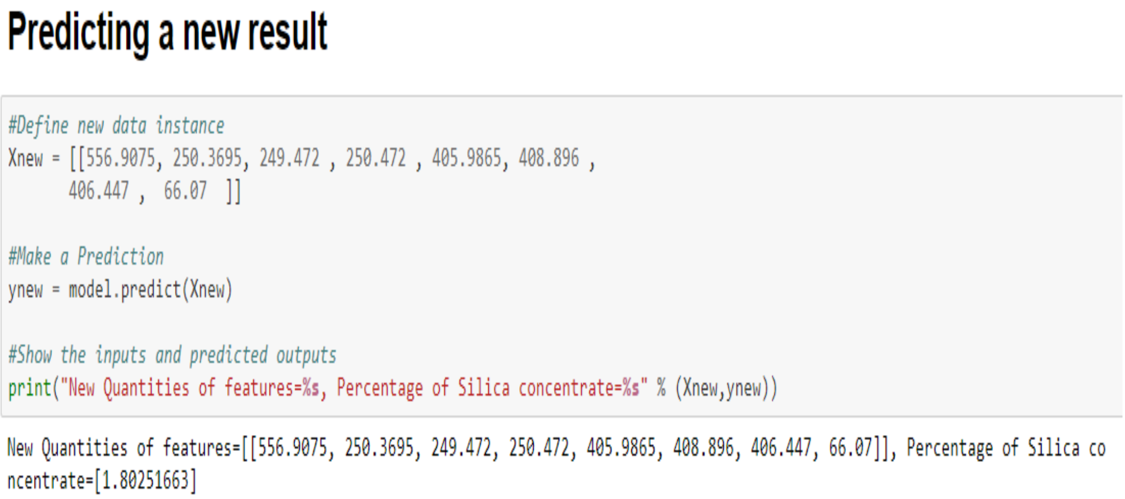
Step8: Predicting Results



```python
predictions = model.predict(X)
```

```python
predictions
```

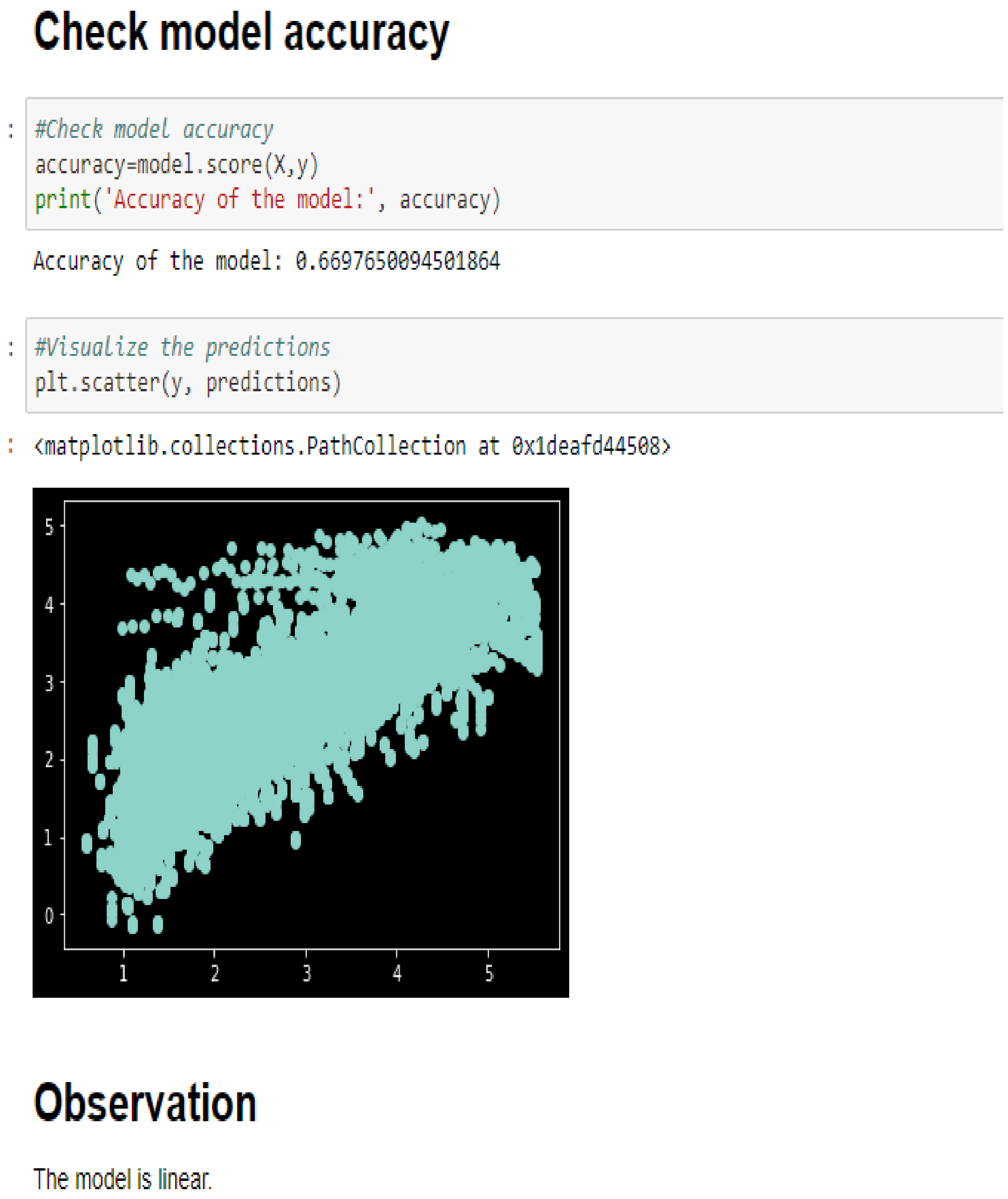
```
array([1.80251663, 1.82270838, 0.791457  , ..., 1.37502984, 3.25750687,
       4.39276254])
```



Predicting a new result

```python
#Define new data instance
Xnew = [[556.9075, 250.3695, 249.472 , 250.472 , 405.9865, 408.896 ,
        406.447 ,  66.07  ]]

#Make a Prediction
ynew = model.predict(Xnew)

#Show the inputs and predicted outputs
print("New Quantities of features=%s, Percentage of Silica concentrate=%s" % (Xnew,ynew))
```

```
New Quantities of features=[[556.9075, 250.3695, 249.472, 250.472, 405.9865, 408.896, 406.447, 66.07]], Percentage of Silica co
ncentrate=[1.80251663]
```



Check model accuracy

```
: #Check model accuracy
  accuracy=model.score(X,y)
  print('Accuracy of the model:', accuracy)

  Accuracy of the model: 0.6697650094501864
```

```
: #Visualize the predictions
  plt.scatter(y, predictions)

: <matplotlib.collections.PathCollection at 0x1deafd44508>
```

Observation

The model is linear.

Fig 3.2.7 Predicting Results

3.3 DATA PREPARATION:

Kaggle is an online community for descriptive analysis and predictive modelling. It collects variety of research fields' dataset from data analytic practitioners. Data scientists compete to build the best model for both descriptive and predictive analytic. It however allows individual to access their dataset in order create models and also work with other data scientist to solve various real-world analytics problems. The input dataset used in developing this model has been downloaded from Kaggle [23]. The dataset contains design characteristics of iron ore froth flotation processing plant which were put together within three (3) months. This is nicely organized using common format and a standardized set of associate features of iron ore froth flotation system.

The dataset contains 24 columns representing the measurements, 737,453 samples exist. The 24 columns include the date and time of the measurement, which will not be used as an input feature. The last columns of the dataset represent the targets of this prediction task.

Description of variables in the dataset:

% Iron Feed: refers to the percentage of iron in the ore entering the flotation process.

% Silica Feed: refers to the percentage of silica in the ore entering the flotation process.

Starch Flow: the flow of the chemical reagent used to make the iron ore decant in the flotation columns.

Amina Flow: the flow of the chemical reagent used to make the silica float in the flotation columns.

Ore Pulp Flow: flow of the ore being fed into the flotation process.

Ore Pulp pH: pH of the ore being fed into the flotation process.

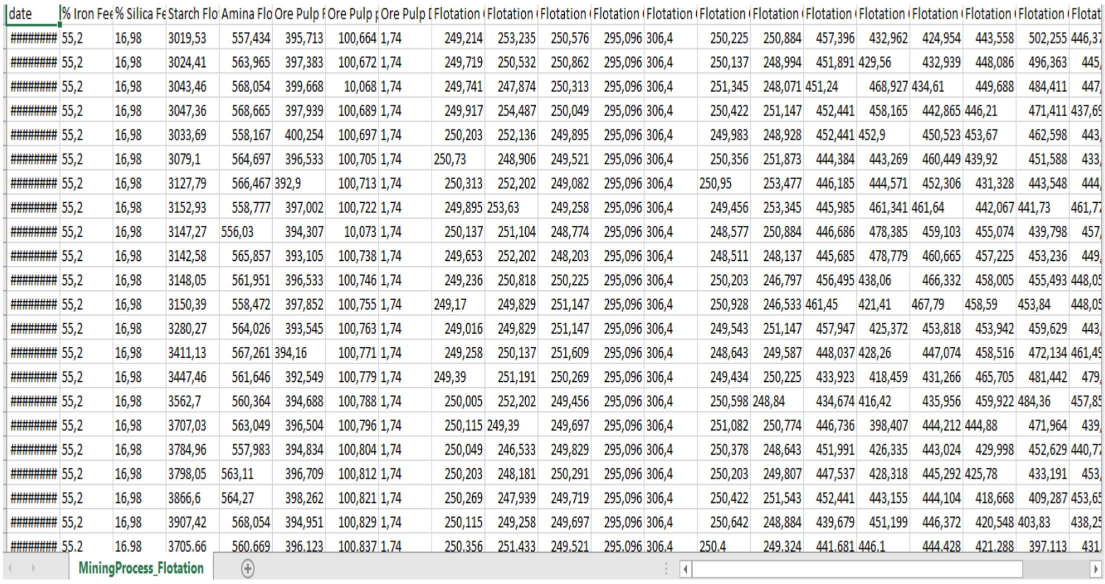
Ore Pulp Density: density of the ore being fed into the flotation process.

Flotation Column Air Flow 1 to 7: operational parameters of the flotation process.

Flotation Column Level 1 to 7: operational parameters of the flotation process.

% Silica Concentrate: the percentage of silica in the concentrated ore obtained at the end of the process. This is the output variable that the model has to predict.

% Iron Concentrate: the percentage of iron in the concentrated ore obtained at the end of the process.



date	% Iron Fee	% Silica Fe	Starch Flo	Amina Flo	Ore Pulp F	Ore Pulp p	Ore Pulp D	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotation (	Flotat
########	55,2	16,98	3019,53	557,434	395,713	100,664	1,74	249,214	253,235	250,576	295,096	306,4	250,225	250,884	457,396	432,962	424,954	443,558	502,255	446,37
########	55,2	16,98	3024,41	563,965	397,383	100,672	1,74	249,719	250,532	250,862	295,096	306,4	250,137	248,994	451,891	429,56	432,939	448,086	496,363	445,
########	55,2	16,98	3043,46	568,054	399,668	10,068	1,74	249,741	247,874	250,313	295,096	306,4	251,345	248,071	451,24	468,927	434,61	449,688	484,411	447,
########	55,2	16,98	3047,36	568,665	397,939	100,689	1,74	249,917	254,487	250,049	295,096	306,4	250,422	251,147	452,441	458,165	442,865	446,21	471,411	437,69
########	55,2	16,98	3033,69	558,167	400,254	100,697	1,74	250,203	252,136	249,895	295,096	306,4	249,983	248,928	452,441	452,9	450,523	453,67	462,598	443,
########	55,2	16,98	3079,1	564,697	396,533	100,705	1,74	250,73	248,906	249,521	295,096	306,4	250,356	251,873	444,384	443,269	460,449	439,92	451,588	433,
########	55,2	16,98	3127,79	566,467	392,9	100,713	1,74	250,313	252,202	249,082	295,096	306,4	250,95	253,477	446,185	444,571	452,306	431,328	443,548	444,
########	55,2	16,98	3152,93	558,777	397,002	100,722	1,74	249,895	253,63	249,258	295,096	306,4	249,456	253,345	445,985	461,341	461,64	442,067	441,73	461,77
########	55,2	16,98	3147,27	556,03	394,307	10,073	1,74	250,137	251,104	248,774	295,096	306,4	248,577	250,884	446,686	478,385	459,103	455,074	439,798	457,
########	55,2	16,98	3142,58	565,857	393,105	100,738	1,74	249,653	252,202	248,203	295,096	306,4	248,511	248,137	445,685	478,779	460,665	457,225	453,236	449,
########	55,2	16,98	3148,05	561,951	396,533	100,746	1,74	249,236	250,818	250,225	295,096	306,4	250,203	246,797	456,495	438,06	466,332	458,005	455,493	448,05
########	55,2	16,98	3150,39	558,472	397,852	100,755	1,74	249,17	249,829	251,147	295,096	306,4	250,928	246,533	461,45	421,41	467,79	458,59	453,84	448,05
########	55,2	16,98	3280,27	564,026	393,545	100,763	1,74	249,016	249,829	251,147	295,096	306,4	249,543	251,147	457,947	425,372	453,818	453,942	459,629	443,
########	55,2	16,98	3411,13	567,261	394,16	100,771	1,74	249,258	250,137	251,609	295,096	306,4	248,643	249,587	448,037	428,26	447,074	458,516	472,134	461,49
########	55,2	16,98	3447,46	561,646	392,549	100,779	1,74	249,39	251,191	250,269	295,096	306,4	249,434	250,225	433,923	418,459	431,266	465,705	481,442	479,
########	55,2	16,98	3562,7	560,364	394,688	100,788	1,74	250,005	252,202	249,456	295,096	306,4	250,598	248,84	434,674	416,42	435,956	459,922	484,36	457,85
########	55,2	16,98	3707,03	563,049	396,504	100,796	1,74	250,115	249,39	249,697	295,096	306,4	251,082	250,774	446,736	398,407	444,212	444,88	471,964	439,
########	55,2	16,98	3784,96	557,983	394,834	100,804	1,74	250,049	246,533	249,829	295,096	306,4	250,378	248,643	451,991	426,335	443,024	429,998	452,629	440,77
########	55,2	16,98	3798,05	563,11	396,709	100,812	1,74	250,203	248,181	250,291	295,096	306,4	250,203	249,807	447,537	428,318	445,292	425,78	433,191	453,
########	55,2	16,98	3866,6	564,27	398,262	100,821	1,74	250,269	247,939	249,719	295,096	306,4	250,422	251,543	452,441	443,155	444,104	418,668	409,287	453,65
########	55,2	16,98	3907,42	568,054	394,951	100,829	1,74	250,115	249,258	249,697	295,096	306,4	250,642	248,884	439,679	451,199	446,372	420,548	403,83	438,25
########	55,2	16,98	3705,66	560,669	396,123	100,837	1,74	250,356	251,433	249,521	295,096	306,4	250,4	249,324	441,681	446,1	444,428	421,288	397,113	431,

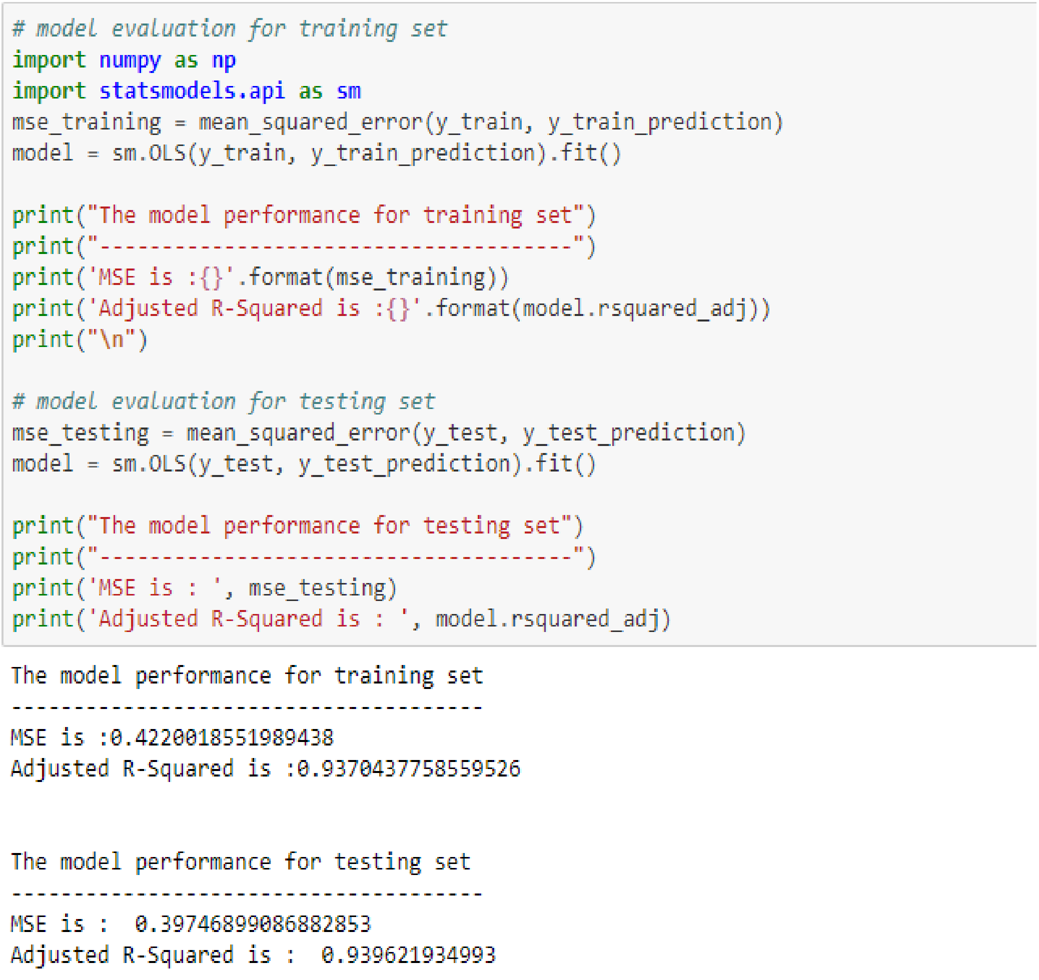
Fig 3.3.1 Dataset Image

CHAPTER 4
RESULTS AND DISCUSION

Experimental Analysis:

In this work we used three regression algorithms, multiple linear regression is used to build the model and after comparing Random Forest, Decision Trees with multiple linear regression it shows best results. And also, we used Deep Learning technique that is Long Short-Term Memory and it has best RMSE and R2 score.

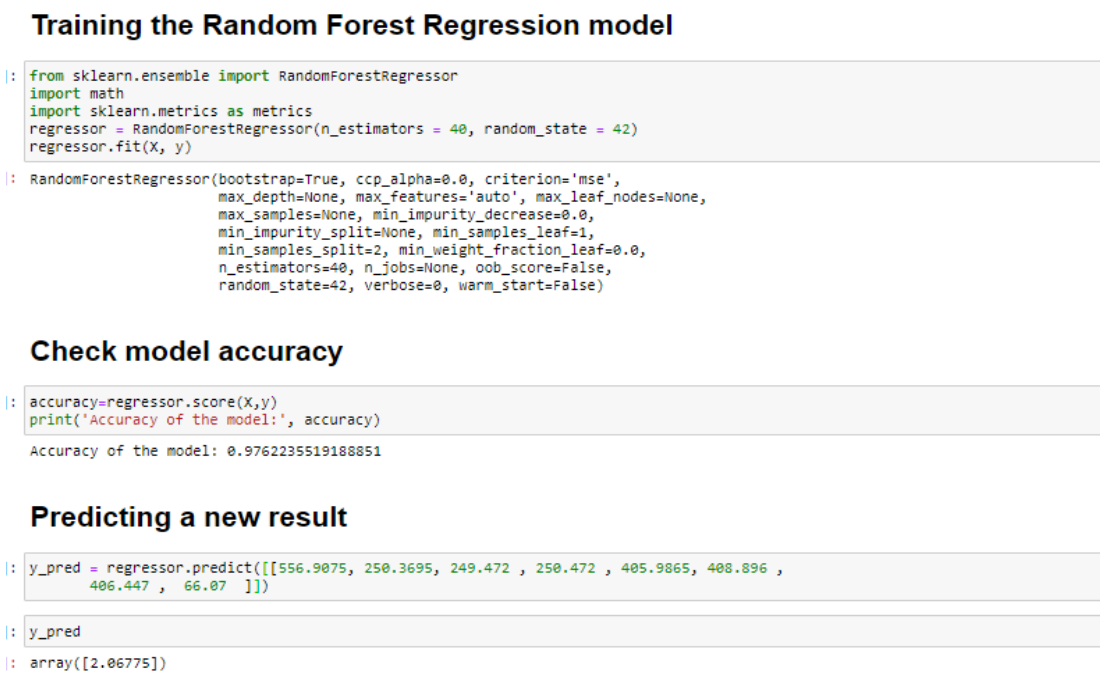
Using Multiple Linear Regression:



```
# model evaluation for training set
import numpy as np
import statsmodels.api as sm
mse_training = mean_squared_error(y_train, y_train_prediction)
model = sm.OLS(y_train, y_train_prediction).fit()

print("The model performance for training set")
print("--------------------------------------")
print('MSE is :{}'.format(mse_training))
print('Adjusted R-Squared is :{}'.format(model.rsquared_adj))
print("\n")

# model evaluation for testing set
mse_testing = mean_squared_error(y_test, y_test_prediction)
model = sm.OLS(y_test, y_test_prediction).fit()

print("The model performance for testing set")
print("--------------------------------------")
print('MSE is : ', mse_testing)
print('Adjusted R-Squared is : ', model.rsquared_adj)
```

```
The model performance for training set
--------------------------------------
MSE is :0.4220018551989438
Adjusted R-Squared is :0.9370437758559526


The model performance for testing set
--------------------------------------
MSE is :  0.39746899086882853
Adjusted R-Squared is :  0.939621934993
```

Fig 4.1.1 MLR Model Evaluation



Training the Random Forest Regression model

```
from sklearn.ensemble import RandomForestRegressor
import math
import sklearn.metrics as metrics
regressor = RandomForestRegressor(n_estimators = 40, random_state = 42)
regressor.fit(X, y)
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=40, n_jobs=None, oob_score=False,
                      random_state=42, verbose=0, warm_start=False)
```

Check model accuracy

```
accuracy=regressor.score(X,y)
print('Accuracy of the model:', accuracy)
```

```
Accuracy of the model: 0.9762235519188851
```

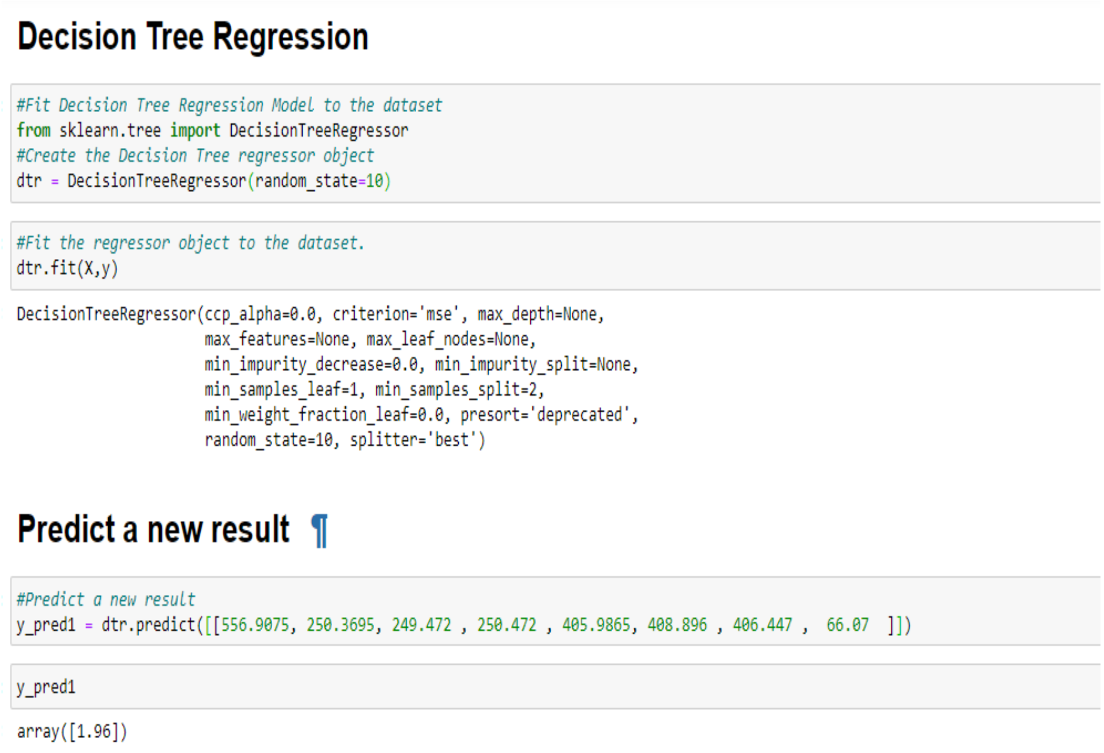
Predicting a new result

```
y_pred = regressor.predict([[556.9075, 250.3695, 249.472 , 250.472 , 405.9865, 408.896 ,
       406.447 ,  66.07   ]])
```

```
y_pred
```

```
array([2.06775])
```

Fig 4.1.2 Predicting using Random Forest



Decision Tree Regression

```
#Fit Decision Tree Regression Model to the dataset
from sklearn.tree import DecisionTreeRegressor
#Create the Decision Tree regressor object
dtr = DecisionTreeRegressor(random_state=10)
```

```
#Fit the regressor object to the dataset.
dtr.fit(X,y)
```

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=None,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=10, splitter='best')
```

Predict a new result ¶

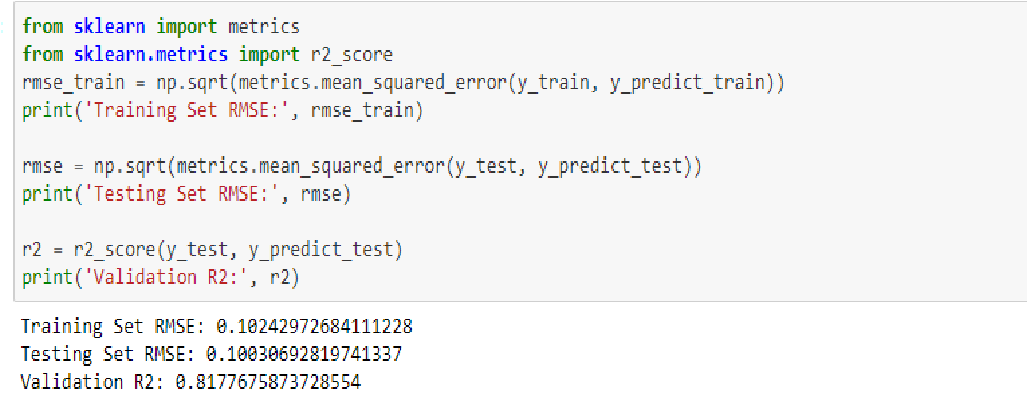
```
#Predict a new result
y_pred1 = dtr.predict([[556.9075, 250.3695, 249.472 , 250.472 , 405.9865, 408.896 , 406.447 ,  66.07   ]])
```

```
y_pred1
```

```
array([1.96])
```

Fig 4.1.3 Predicting using Decision Tree

Using Long Short-Term Memory:



```python
from sklearn import metrics
from sklearn.metrics import r2_score
rmse_train = np.sqrt(metrics.mean_squared_error(y_train, y_predict_train))
print('Training Set RMSE:', rmse_train)

rmse = np.sqrt(metrics.mean_squared_error(y_test, y_predict_test))
print('Testing Set RMSE:', rmse)

r2 = r2_score(y_test, y_predict_test)
print('Validation R2:', r2)
```

```
Training Set RMSE: 0.10242972684111228
Testing Set RMSE: 0.10030692819741337
Validation R2: 0.8177675873728554
```

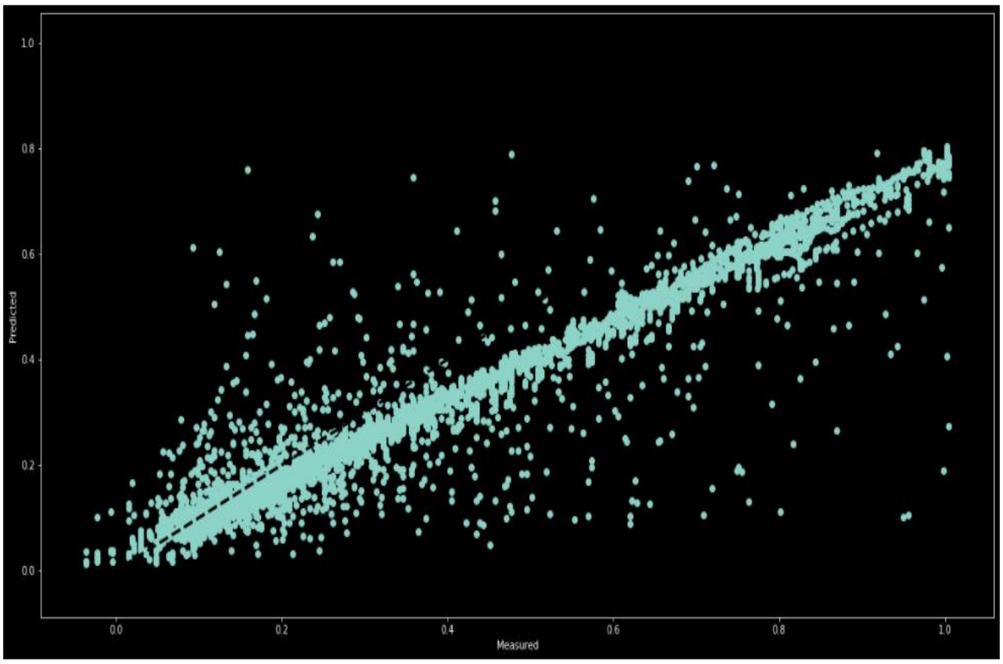
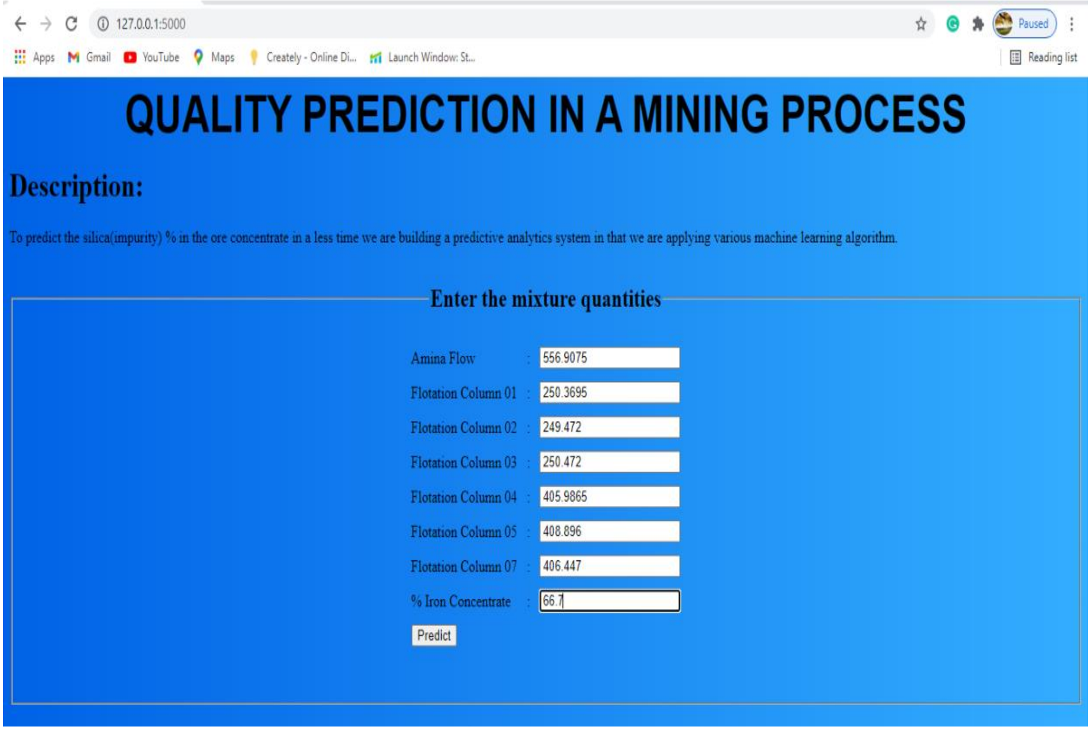


Fig 4.1.4 LSTM Model Evaluation

Results:

The output is seen through user interface which it consists the fields to upload the values of various mixtures and it shows the output by considering the input as uploaded values that and after clicking predict button it gives the percentage of silica impurity as show in below user interfaces.



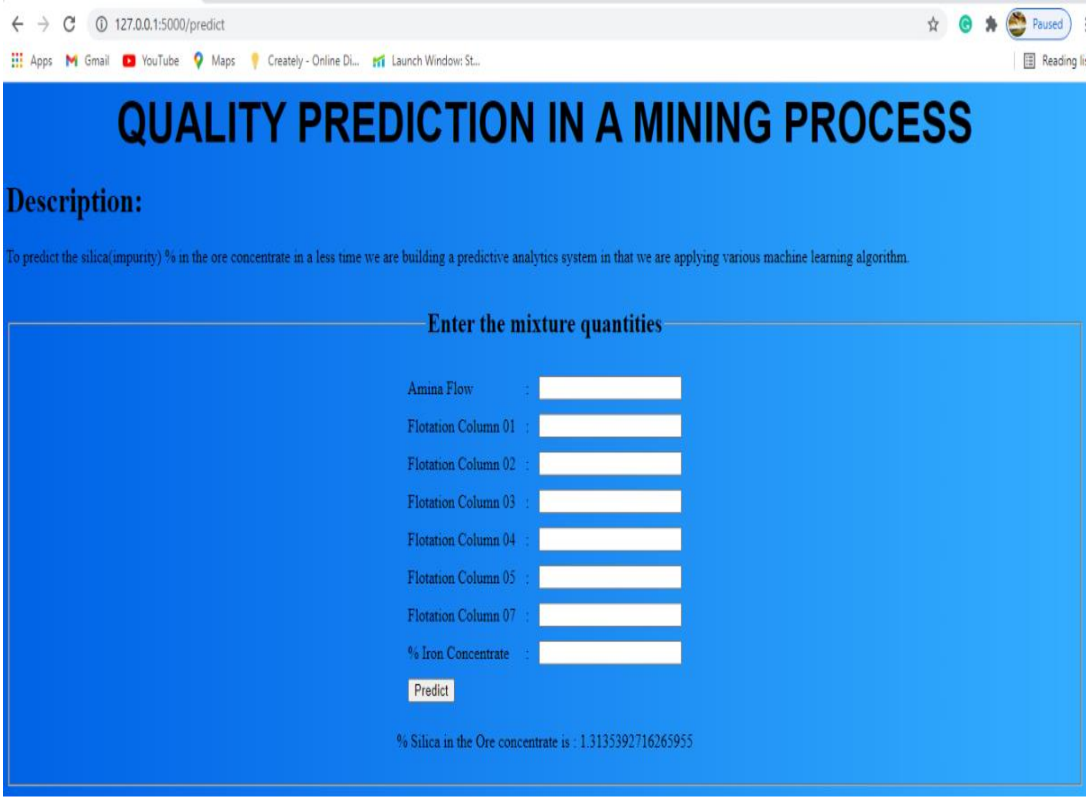
QUALITY PREDICTION IN A MINING PROCESS

Description:

To predict the silica(impurity) % in the ore concentrate in a less time we are building a predictive analytics system in that we are applying various machine learning algorithm.

Enter the mixture quantities

Amina Flow	:	556.9075
Flotation Column 01	:	250.3695
Flotation Column 02	:	249.472
Flotation Column 03	:	250.472
Flotation Column 04	:	405.9865
Flotation Column 05	:	408.896
Flotation Column 07	:	406.447
% Iron Concentrate	:	66.7



QUALITY PREDICTION IN A MINING PROCESS

Description:

To predict the silica(impurity) % in the ore concentrate in a less time we are building a predictive analytics system in that we are applying various machine learning algorithm.

Enter the mixture quantities

Amina Flow	:	
Flotation Column 01	:	
Flotation Column 02	:	
Flotation Column 03	:	
Flotation Column 04	:	
Flotation Column 05	:	
Flotation Column 07	:	
% Iron Concentrate	:	

% Silica in the Ore concentrate is : 1.3135392716265955

Fig 4.1.5 Predicted output

DISCUSSION:

- The findings from this study suggest that machine learning algorithms have the predictive power to predict percentage of silica concentrate in iron ore froth flotation processing plant in real-time as opposed to 2 hours laboratory analysis.
- However, after the 3 months' observations of iron ore froth flotation processing plant dataset was analyzed, on average, the silica concentrate predictions will be off by 0.38% with a standard deviation of approximately 0.12%, which is significant considering the fact that silica concentrate ranges from 0.77% to 5.53%.
- This result should be interpreted with caution because the silica concentrate variable used in the analysis was lagged 2 hours and could be further explored with diverse residence time.
- However, it is worth noting that each observation in the froth flotation plant can be estimated with respect to silica concentrate as fast as possible. This further connotes that not only the execution time is significant but also the precision of the predictive task.
- Thus, when both the prediction accuracy and execution time are significant features of an automating the froth flotation plant system, the best option is Recurrent Neural Networks like LSTM. This provides in effective predictive in real-time.

CHAPTER -5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION:

The present study used to predict percentage of silica concentrate in iron ore froth flotation processing plant in real-time as opposed to 2 hours laboratory analysis. However, by implementing Machine Learning Model, know we can predict the percentage of silica in ore concentration without any delay of time and help the engineers, giving them early information to take actions and reduce the impurities. And by developing flask web application in which we can provide a user interface in which engineers can have the values earlier and they can predict the target variable i.e., %Silica Concentrate by having different values in the dataset.

The findings from this study suggest that machine learning algorithms have the predictive power to predict percentage of silica concentrate in iron ore froth flotation processing plant in real-time as opposed to 2 hours laboratory analysis. However, after the 3 months' observations of iron ore froth flotation processing plant dataset was analyzed, on average, the silica concentrate predictions will be off by 0.38% with a standard deviation of approximately 0.12%, which is significant considering the fact that silica concentrate ranges from 0.77% to 5.53%.

5.2 FUTURE SCOPE:

The dataset analyzed in this study was small and we have scope to deal with large datasets by using different techniques in Machine Learning and Deep Learning. And also, we can collect different datasets across the world and deal with them and know the best results to predict different impurities in ore concentration. On the other hand, we can extend the application of the methodology for different froth flotation processing plants preferably paper mills industry and mineral processing.

BIBLIOGRAPHY

[1] Dawson, P., & Koorts, R. (2014). Flotation Control Incorporating Fuzzy Logic and Image Analysis: IFAC Proceedings Volumes, 47(3), 352–357.

[2] Luo, X., Feng, B., Wong, C., Miao, J., Ma, B., & Zhou, H. (2016). The critical importance of pulp concentration on the flotation of galena from a low-grade lead–zinc ore. Journal of Materials Research and Technology, 5(2), 131–135.

[3] Horn, Z. C., Auret, L., McCoy, J. T., Aldrich, C., & Herbst, B. M. (2017). Performance of Convolutional Neural Networks for Feature Extraction in Froth Flotation Sensing. IFACPapersOnLine, 50(2), 13–18.

[4] Nakhaei, F., & Irannajad, M. (2015). Application and comparison of RNN, RBFNN and MNLR approaches on prediction of flotation column performance. International Journal of Mining Science and Technology, 25(6), 983–990.

[5] Bergh, L. G., & Yianatos, J. B. (2011). The long way toward multivariate predictive control of flotation processes. Journal of Process Control, 21(2), 226–234.

[6] Eduardo Oliveira, Pablo Drummond, Daniele Kappes, Cássio de Moraes and Mariana Teixeira.

(2018). Traditional Machine Learning or Deep Learning, in Conference 22° Seminário de Automação e TI, , São Paulo, Brazil.

GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.

PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Project Outcomes	
Course Outcome (CO1)	Identify and analyze the problem statement using prior technical knowledge in the domain of interest.
Course Outcome (CO2)	Design and develop engineering solutions to complex problems by employing systematic approach.
Course Outcome (CO3)	Examine ethical, environmental, legal and security issues during project implementation.
Course Outcome (CO4)	Prepare and present technical reports by utilizing different visualization tools and evaluation metrics.

Mapping Table

CS1537 : MAIN PROJECT															
Course Outcomes	Program Outcomes and Program Specific Outcome														
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	3	3	1					2	2	2				1	1
CO2	3	3	3	3	3			2	2	2		1		3	3
CO3	2	2	3	2	2	3	3	3	2	2	2			3	
CO4	2		1		3				3	3	2	2		2	2

Note: Map each project outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1- Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped

QUALITY PREDICTION IN A MINING PROCESS

P. Praveen, N. Chandrika, P. Vignesh, M. Srikanth, student CSE Department & GEC, Gudlavalleru, Krishna
Mr. M. N. Satish Kumar Assistant Prof. CSE Department & GEC, Gudlavalleru, Krishna
maganti.nagasatishkumar@gmail.com

Abstract

The% of Silica is measured in a lab experiment it takes at least one hour for the process engineers to have this value. As this impurity is measured every hour and it takes a lot of time for a day and causes delay in the mining process. The environment is polluting while reducing the number of ore that goes to tailings as you reduce silica in the ore concentrate. The overall goal is to predict impurity in the ore concentrate in mining process. In this case impurity is specifically Silica concentrate. Silica concentrate is a measured variable but takes time to report results, thus reducing efficiency in the mining process. Being able to predict the silica content without stopping to test is the extended goal of this project. This appears to be a continuous batch process, where raw material is fed into a flotation system, processed, removed, and the process repeated. The purpose is to evaluate the feasibility of using machine learning algorithms like Multiple Linear Regression, Random Forest and Decision tree to predict in real-time. And also, by using Deep Learning techniques like LSTM, we can predict the silica impurity in the ore in less time and help the engineers for early prediction and reduce the impurities. We also developed a web application to display the prediction. The web application is built by using flask framework and it is integrated with trained ML model and it 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.

Keywords— *Mining Process,* Machine Learning, Deep Learning, LSTM.

I. Introduction

Data storage in various format such as records, files, documents, sound, movies, science and a lot of new information forms has been driven by the emergence of information technology in numerous domains For better decision-making, the data generated from diverse applications require an appropriate way of extracting information from big repositories. The purpose is to discover usable information from the wide collection of data in databases (KDD), often dubbed data extraction (Data Mining). In order to find and extract patterns of recorded data, the basic features of data mining are numerous approaches and algorithms. Data mining and application for knowledge discovery have become an important component in many organizations, as they play an important part in decision making. In the new sectors of statistics, databases, machine learning, model reorganization and artificial intelligence, computer capabilities etc., data mining technology has been incorporated.

In several industries, a model for quality prediction was established for the production of faultless products. However, in single-stage production, most quality prediction model is established. Previous research demonstrate that one-stage quality system in multi-stage production cannot effectively tackle the quality problem.

Linear regression is a statistical study that determines how a relationship between two types of variables is modeled, dependent and independent (predictor). Regression has major goal to investigate if independent factors have predicted the result variable well and which independent variables are important predictors of the result.

There has recently been increasing interest in investigating primordial factors for iron ore recovery at a froth flotation treatment plant. The Financial Times shows that ore iron is more of a raw material for steel manufacture than any other commodity, with the exception perhaps of land, in the world economy. Every tonne of iron ores produced has been shown to discharge around 2.5-3.0 tonnes. In addition, figures indicate that about 130 million tonnes of iron ore are produced annually. This suggests that if the mining reservoirs contain, for instance, an average of around 12% iron ore, over

1.52 million tonnes of iron would be waste per year. In the brotherhood of iron mining, stakeholders rely on standard laboratory testing techniques, which generally take more than two hours to attain their target quality in the froth flotation processing facility. Since environmental protection is highly important and good iron grade is needed which is mined from ore. Then we may forecast a single dependent variable with two or three separate versions by applying machine learning methods such as Multiple Linear Regression, and we can employ Random Forest Decision Trees further. Deep learning techniques such as LSTM's neural network are utilized to predict silica impurity in mining process, which is well known for its time series prediction applications.

1.1 Data Mining

Data mining is used to use massive amounts of data to find hidden patterns and associations that are useful in decision-making. Alternatively, exploratory analysis, discovery by data and inferior learning were called. In this standard access, data mining access to a database differs in a number of ways: query, data and output. A data mining algorithm is a well-defined technique in which data is taken as input and generated as models or patterns. The phrase well-defined shows that it is precisely possible to encode the operation as a certain number of rules. In order to characterize the whole (most of the) data set, the structures found during the data mining process are called "models." There are also occasions where the identified structures have some local data characteristics and the term pattern is applied in this case.

1.2 Considerations for applying data mining

In order to construct an effective data mining solution, the user has to investigate and articulate his objective. The problem objective leads the user to the correct learning algorithm. When hidden groupings in data may be detected or a connection between key data variables is established, users want to find information and select a technique for clustering or the association mining. Alternatively, a predictive model may be created that may divide samples into a category such as low air quality or a real world result such as an aviation quality score. There are a big, rising number of algorithms inside the prediction paradigm and the knowledge finding paradigm. The choice between the methods of any paradigm is a problem of his own accord. Domingos highlights some of the important issues in helping new practitioners in implementing algorithms for machine learning. When making this decision, the User should consider the intricacy and quantity of data presented. For example, a basic linear classifier is not suitable for a sophisticated, non-linear classification task. However it requires the employment of advanced study methods, such as deep artificial neural networks, in consideration of considerations regarding the storage, memory and durations of training.

1.2.1 Predicting quality

Quality prediction involves the development of models in which quality input features are related to quality inputs and the use of models in order to forecast what the resulting quality property value will be of a collection of input parameters. For predication, regression approach can be adjusted. The regression analysis can be employed for modeling a link between one or more separate and dependent variables. Individual variables are already known attributes for data mining, and the answer variables are to be anticipated. Sadly, not just predictions are numerous real-world difficulties. Therefore, the prediction of future values can require more advanced algorithms (e.g. regression of logistics, decision trees or neural nets). For regression and classification, the same model types can often be utilized.

1.3 Neural networks

The neural grid is a set of interconnected I/O modules with a connecting weight. The weight change of the network will enable the appropriate input to be anticipated during the learning phase. Neural networks can be used to uncover patterns and to find trends that human or computer technology can recognize far too difficult to draw significance from concatenated or inaccurate data. They are perfect for inputs and outputs that are valued continuously. Neural networks are ideally suited for prediction or forecasting requirements when determining data patterns or trends.

II. Literature Review

Marco Canaparo et. all (2019) In this study, the data mining strategies were initially comparable as far as software fault prognosis was concerned. The author employed existing literature to collect online data set, procedures and performance criteria in order to attain this objective. authors paid greater attention to open source and deep learning approaches than earlier studies. Data set linked to open source projects. By analyzing the findings, the author can find the best average accuracy of all data sets achieved by Bagging and Random Forest. Data mining can also serve to determine and predict software quality and can be used in conjunction with statistical analysis. [1]

Brijesh Kumar Baradwaj et. all (2011) This research uses classification tasks to forecast the division of students on the basis of an old database from a student database. Since there are numerous ways for the classification of data, the decision-tab method is being applied. Information like attendance, class testing, seminar and markings have been collected from the previous database of the student to determine performance at the end of the semester. The students and teachers can raise the division of students through this research. This study will also highlight those students who have to pay special attention to lesser degrees of failure and take appropriate steps for the upcoming semester. [2]

Yunus Koloğlu et. all (2012) The successful and invaluable players gather economically in their clubs and generally young talents such as Kylian Mbabpe and Paulo Dybala are worthwhile. 180 and 184 have proven beneficial, assuming that larger players do not have dribble skills and that shorter players lack air ball control. The premier league is also recognized as the most difficult league in Europe as another thumb rule, therefore there is no surprise that English players are more valuable. The fact that card numbers have not affected the player's market value is only an interesting feature in the study, which can be explained by the fact that valued players are more careful to avoid a charge. In general, a more reasonable data collection might be used to improve the study. [3]

Thuraiya Mohd et. all (2020) In this research, a qualitative and quantitative factors (dumb variables) were empirically experimented utilizing the property dataset in the Kuala Lumpur area, Malaysia. The results revealed that statistically significant contributions have been achieved by elements like the main floor area, Green Certification, Tenure and Number of Bedrooms. In other words, all these factors played key roles in the prices of transactions. Main floor area characteristics, green certificate and tenure are related to the transaction price positively. The main floor area has provided the most contribution to the model based on the standardized beta coefficient. [4]

Rajat Chaudhari et. all (2020) After an analysis of a variety of documents, author find that the soil fertility forecast will help to decrease farmers' troubles and to offer farmers with effective information to achieve high yields and hence maximize earnings, thereby reducing suicide rates and reducing their problems. To predict soil fertility, a model is implemented. The system uses supervised and uncontrolled algorithms to learn the machines and provides the highest possible precision results. It compares results from the four algorithms and selects the one that gives the best and most precise output. [5]

Turóczy Zsuzsannaa et. all (2012) The major objective is to enhance the competitiveness, flexibility, adaptability and reactivity of ceramic companies. Since the ceramic sector is an essential part of manufacturing, authors concentrated on this area in order to assess the progress of companies in the sector. The value of the research lies in its novelty and efficacy, as in the instance of a company producing advanced ceramic items, the performance indicators have been analyzed by multiple regression analyses. This analysis is frequently a multivariate and explanatory approach of analysis. Regression analysis describes the connection between a dependent variable and several distinct factors. [6]

Fahmi Arif et. all (2013) In this study, algorithms of decision tree are utilized to disclose the relationship between factors of product and the final product quality level. The consistent number of values for all product characteristics as the attribute of the decision tree is very low, low, low medium, upper medium, high and very high, all of which are extremely low. It is also possible to summarize that ID3 performs better than C4.5 and CHAID in the case of imbalanced datasets with a uniform number of attribute values, whereas DS, RT and RF fail in the classification of minorities. [7]

E. V. Ramana et. all (2017) Neural net and rule induction models were surpassed by a 95 percent prediction accuracy on the test data set over Naive Bayes model (80 percent). The rule induction model shows that sink marks are created by high temperature of moulding, low injection velocity, dust temperature and injection duration. [8]

T B Chistyakova et. all (2019) A computer data mining approach is described to forecast quality in multi-sortiment, large-scale and multinational polymer films. The following paper provides a library of statistical and data mining methods that enable the testing of normal distribution data, the predicting of the quality of film polymers for various line configurations and film types using statistical tests. The methods include recurring neural networks, a long-term memory neural network and a convolutions network. [9]

Umesh Kumar Pandey et. all (2011) The student database uses the Bayesian classification method to forecast the division of students on the basis of the preceding year record. This research helps students and teachers increase the student's division. The study will also identify pupils who needed additional attention to reduce failures and take suitable action in due course. [10]

Amjad Abu Saa et. all (2016) Multiple data mining tasks were performed in this research article to develop qualitative prediction models that could efficiently and efficiently predict student grades from a dataset of collected training. The first was a survey that targeted and collected a number of personal, social and academic information about university students. Secondly, the obtained data set has been pre-processed and examined so that data mining jobs are suitable. Thirdly, data mining operations were carried out on the dataset in hand to create and test categorization models. [11]

Colin Bellinger et. all (2017) The number of articles reporting on the use of data mining tools to monitor air pollution has increased considerably from our survey. This is because massive data groups and computer power are being made more available and because authors are becoming aware of the potential advantages of data mining. Despite this tendency and the possible advantages within the sector, a study of the current state-of-the art has not been carried out to the best of our knowledge. [12]

III. Methodology

Gathering the data is the first step to build a model and the process of gathering data depends on the type of project and how we are collected or taken the data. Kaggle is one of the most popular websites where we can download free dataset which is related to the project and we downloaded the dataset that contains 24 features like %iron feed, %silica feed, starch flow, Amina flow, Ore Pulp pH, Ore Pulp Density, Flotation Column 01- 07 Air Flow, Flotation Column 01-07, % Iron Concentrate, % Silica Concentrate and737453 samples which include data and time. Data Preprocessing is done for cleaning the data and removing the out layers which are present in the dataset. In data preprocessing step outliers are removed by IQR score method. After that by using matplotlib we can visualization the data and analyze the data. For building a regression model this are the five steps in the following figure 3.1.

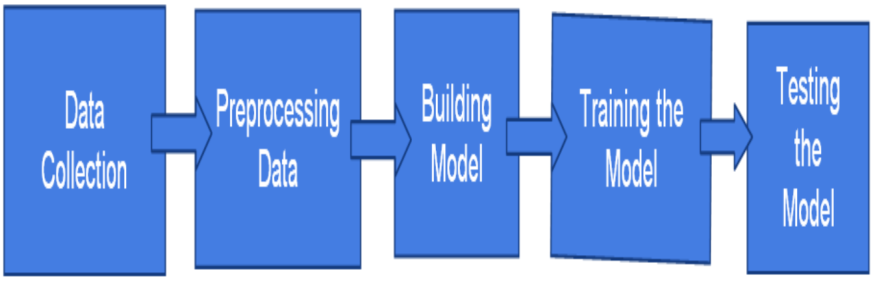


Figure: 1 Building ML Model

Data Collection

Data collection is defined as the process of accurate research insights using standard verified methodologies to be collected, measured and analyzed. On the basis of acquired data, a researcher might evaluate his theory. In most circumstances, data gathering, irrespective of the subject of research, is the main and most significant step in research.

Preprocessing Data: Data Preprocessing is the step of any machine learning process in which the data is modified or encoded to make it so easy to comprehend. In other words, the data's characteristics may now easily be read using the algorithm.

Building Model: Construction of the model Regression analysis involves the development of a probabilistic model, in which the link between the dependent and the independent variables is better described. The multiple linear regression attempts by fitting a linear equation into the observe data to modeled the association between the two or more explicatory variable and the response variable. Each value of the independent variable x is linked to the dependent variable y value.

Y = b0 + b1 * x1 + b2 * x2 + b3 * x3 +……bn * xn;
Dependent variable = Y
Independent variables=x1,x2,x3,…..xn

For building Multiple Linear Regression model we need to import the data to the operating environment like jupyter notebook. For loading the data sets and to preprocess them we need to import the libraries such as pandas, numpy and matplotlib for visualization. Check if the dataset consists of any missing values and remove the poorly correlated independent variables by using the correlation heatmap. And then we have split the dataset as 80% of the data to train the model and 20% to test the model. After splitting of data, we can train and test the model.

The following step is to analyze model performance when a machine learning model is built and to comprehend the model that is best. Root Mean Square is the measure of how well a regression line fits the data points. And Adjusted R-squared is used to determine the goodness of fit in regression analysis.

Random Forest is a popular learning system that is frequent in numerous academic publications, competition for the Kaggle and blog posts for categorization purposes. Random forests may also be utilized for regression tasks in addition to classification. The non-linear character of a Random Forest is a fantastic option for it over linear methods. Tree decision also has an excellent regression problem approach and is utilized to forecast silica impurity in an ear in these controlled Machine Learning techniques.

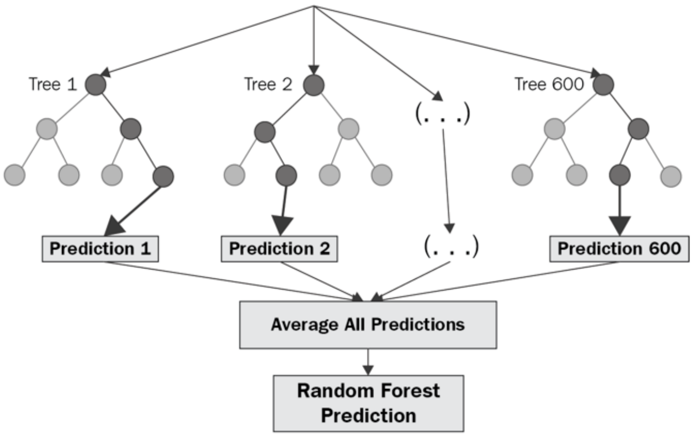


Figure: 2 Random Forest Algorithm

The method of research is based on an examination of regression. This form of analysis is utilized for several variables to be modeled and analyzed. The multiple regression analysis includes a description of the connection between a dependent variable and numerous independent variables.

Then, we can broaden this model by employing profound learning approaches and various ways of using predictive analysis systems, which are a sort of recurring neural network capable of learning orders in sequence prediction issues using Long Short term Memory.

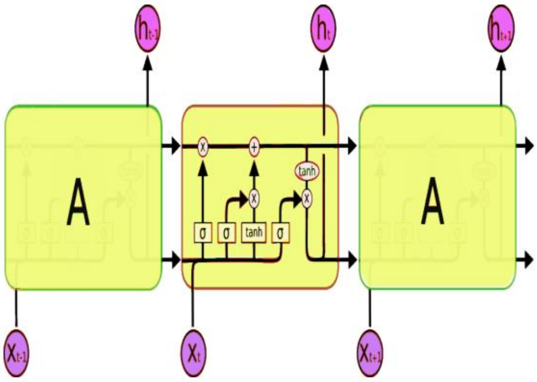
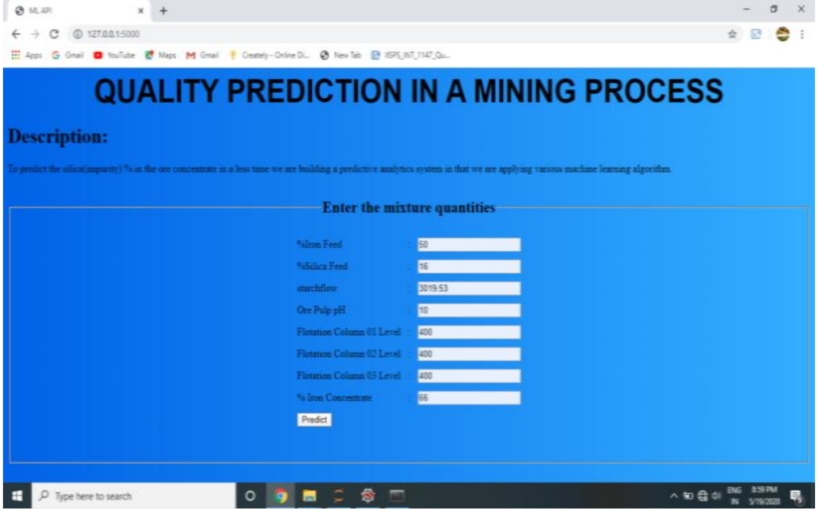


Figure: 3 LSTM Algorithm

IV. Results

After successful trained and tested the models which are build by different algorithms like Multiple Linear Regression, Random Forest, Decision Tress and all are given good results. The results of Multiple Linear Regression are acceptable and we used Multiple Linear Regression to build flask web application and predict the results as shown in the figure and also, we further implemented a LSTM model which gives better results.



QUALITY PREDICTION IN A MINING PROCESS

Description:

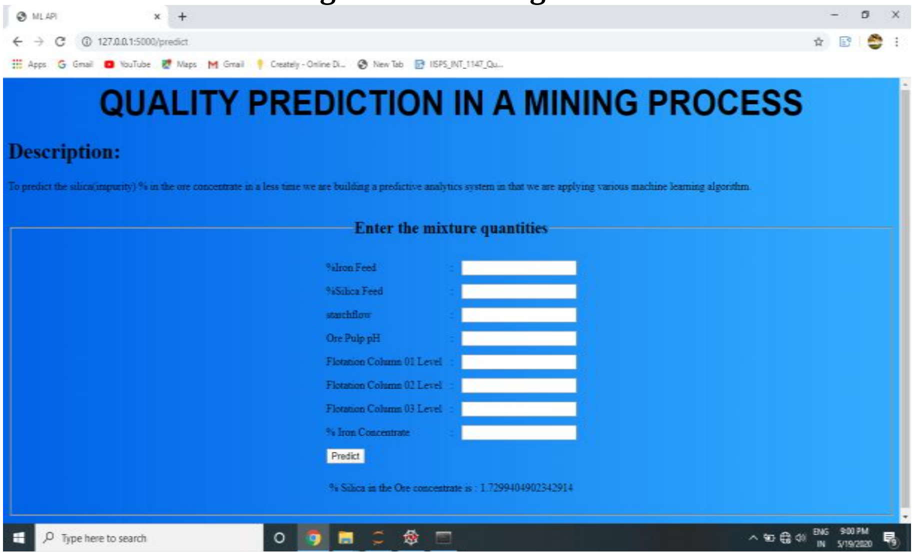
To predict the silica(impurity) % in the ore concentrate in a less time we are building a predictive analytics system in that we are applying various machine learning algorithm.

Enter the mixture quantities

%Iron Feed : 50
%Silica Feed : 16
starchflow : 3019.53
Ore Pulp pH : 10
Flotation Column 01 Level : 400
Flotation Column 02 Level : 400
Flotation Column 03 Level : 400
% Iron Concentrate : 66

Predict

Figure: 4 Entering Data



QUALITY PREDICTION IN A MINING PROCESS

Description:

To predict the silica(impurity) % in the ore concentrate in a less time we are building a predictive analytics system in that we are applying various machine learning algorithm.

Enter the mixture quantities

%Iron Feed :
%Silica Feed :
starchflow :
Ore Pulp pH :
Flotation Column 01 Level :
Flotation Column 02 Level :
Flotation Column 03 Level :
% Iron Concentrate :

Predict

% Silica in the Ore concentrate is : 1.7299404902342914

Figure: 5 Output Result

V. CONCLUSION

The findings from this study suggest that machine learning algorithms have the predictive power to predict percentage of silica concentrate in iron ore froth flotation processing plant in real-time as opposed to 2 hours laboratory analysis. However, after the 3 months' observations of iron ore froth flotation processing plant dataset was analyzed, on average, the silica concentrate predictions will be off by 0.38% with a standard deviation of approximately 0.12%, which is significant considering the fact that silica concentrate ranges from 0.77% to 5.53%.This result should be interpreted with caution because the silica concentrate variable used in the analysis was lagged 2 hours and could be further explored with diverse residence time.

However, it is worth noting that each observation in the froth flotation plant can be estimated with respect to silica concentrate as fast as possible. This further connotes that not only the execution time is significant but also the precision of the predictive task.

Thus, when both the prediction accuracy and execution time are significant features of an automating the froth flotation plant system, the best option is artificial neural network. This provides in effective predictive in real-time.

VI. Future Scope

The dataset analyzed in this study was small and we have scope to deal with large datasets by using different techniques in Machine Learning and Deep Learning. And also, we can collect different datasets across the world and deal with them and know the best results to predict different impurities in ore concentration. On the other hand, we can extend the application of the methodology for different froth flotation processing plants preferably paper mills industry and mineral processing.

REFERENCES

[1.] Marco Canaparo and Elisabetta Ronchieri "Data Mining Techniques for Software Quality Prediction in Open Source Software" EPJ Web of Conferences 2019.

[2.]Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students" Performance" IJACSA 2011.

[3.]Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz "A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position" ABDULLAH GÜL UNIVERSITY INDUSTRIAL ENGINEERING DEPARTMENT 2012.

[4.]Thuraiya Mohd, Syafiqah Jamil and Suraya Masrom "Multiple Linear Regression on Building Price Prediction with Green Building Determinant" International Journal of Advanced Science and Technology 2020.

[5.]Rajat Chaudhari, Saurabh Chaudhari, Atik Shaikh, Ragini Chiloba, Prof.T.D.Khadtare "Soil Fertility Prediction Using Data Mining Techniques" International Journal of Future Generation Communication and Networking 2020.

[6.]Turóczy Zsuzsannaa, Liviu Marian "Multiple regression analysis of performance indicators in the ceramic industry" Emerging Markets Queries in Finance and Business 2012.

[7.]Fahmi Arif, Nanna Suryana, Burairah Hussin "A Data Mining Approach for Developing Quality Prediction Model in Multi-Stage Manufacturing" International Journal of Computer Applications 2013.

[8.]E. V. Ramana, S. Sapthagiri and P. Srinivas "Data Mining Approach for Quality Prediction and Control of Injection Molding Process" Indian Journal of Science and Technology 2017.

[9.]T B Chistyakova and M A Teterin "Data mining system for predicting quality of polymeric films" AMCSM 2020.

[10.] Umesh Kumar Pandey S. Pal "Data Mining : A prediction of performer or underperformer using classification" IJCSIT 2011.

[11.] Amjad Abu Saa "Educational Data Mining & Students' Performance Prediction" IJACSA 2016.

[12.] Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane and Alvaro Osornio-Vargas "A systematic review of data mining and machine learning for air pollution epidemiology" BMC Public Health 2017.